TECHNICAL BULLETIN NO. 210 FEBRUARY, 1931

CORRELATION ALINEMENT CHARTS IN FOREST RESEARCH

A METHOD OF SOLVING PROBLEMS IN CURVILINEAR MULTIPLE CORRELATION

By

DONALD BRUCE
Senior Silviculturist
and
L. H. REINEKE
Assistant Silviculturist, Branch of Research
Forest Service

UNITED STATES DEPARTMENT OF AGRICULTURE, WASHINGTON, D. C.

TECHNICAL BULLETIN No. 210 FEBRUARY, 1931

UNITED STATES DEPARTMENT OF AGRICULTURE
WASHINGTON, D. C.

CORRELATION ALINEMENT CHARTS IN FOREST RESEARCH: A METHOD OF SOLVING PROBLEMS IN CURVILINEAR MULTIPLE CORRELATION

By DONALD BRUCE, *Senior Silviculturist*, and L. H. REINEKE, *Assistant Silviculturist, Branch of Research, Forest Service*

CONTENTS

ADVANTAGES AND LIMITATIONS OF THE GRAPHIC METHOD

Problems in forest research almost inevitably involve a consideration of the relations between two or more variable factors. Results depend, for example, on the increase of tree diameter or volume with age, on the influence of site, on the vitality of seed, or on other similar relations. Although the relation is sometimes one involving cause and effect, this is by no means essential. The interrelation between height and diameter may be of importance although neither is the cause of the other, the two being more or less closely interrelated because they result from causes which are in part common to both. Often, moreover, the investigation is not restricted to a pair of variables but concerns the association of one with two or more others, as in preparing yield tables where the relation is sought between both age and site and such factors as volume per acre. In some cases there is no certain advance knowledge as to what variables are involved, and the first step in the study must be to ascertain what factors are related and what are not.

The graphic method has conventionally been used in solving such problems. It has many advantages, chief among which are its celerity and flexibility. Unfortunately, it lends itself to careless technic and so has led to many false conclusions. This has been caused in part by inherent weaknesses in the method and in part by an inadequate analysis of potential sources of error, together with a poorly developed technic for checking the results obtained. There is need for further development along quantitative lines.

The modern statistical method supplies a powerful and delicate machinery to replace or supplement the simple graphs which foresters have ordinarily used. Its drawbacks are greater intricacy and apparently greater laboriousness. However, its apparent intricacy disappears with use, and the labor involved is more formidable in appearance than in reality. The greater delicacy of the method usually permits a given degree of accuracy to be obtained with a much smaller number of data, providing they are accurate; and in many cases the saving of time in collecting data will more than offset any increased labor in analyzing them. Until recently a third drawback was the rigidity of the method. Its usefulness was restricted to those cases where, graphically, a straight line might be used without serious error or (with a large increase in the labor involved) to those where the underlying curve was of a type for which the form of equation was known. Recently, however, a new statistical method (*9*),[1] [2] partially graphic in its technic, has been devised which is free from this disadvantage and which appears particularly suited to forestry problems. It is this method which will be described in the following pages.[3] Before discussing it, it will be necessary to explain the criteria by means of which the adequacy of any method may be judged, for only in this way can the advantages of the new technic be fully appreciated. By means of this discussion, moreover, it will be possible to define and explain the statistical conceptions used in terms of the graphs with which foresters are familiar, instead of on a purely mathematical basis.

[1] Italic numbers in parentheses refer to the Partial List of Statistical Literature, p. 85.

[2] This new method was originated by Mordecai Ezekiel, Bureau of Agricultural Economics. As herein presented it has been materially modified. These modifications include the use of alinement charts, the investigation of a possible functional form of the dependent variable, and a new type of final correction. These changes considerably reduce the labor involved and increase flexibility.

[3] The authors acknowledge the very valuable assistance in the preparation of the material here presented, rendered by Misses Mary L. Denoyer, Lucile L. Gumaer, and Theresa Hoerner, statistical clerks in the Office of Forest Measurements.

STATISTICAL MEASURES

THE STANDARD ERROR AS A MEASURE OF CURVE ACCURACY

One of the primary purposes of any curve is to permit estimation of values of one factor from given values of another. It is useful perhaps to be able to estimate the height of a tree if its age is known, or the number of trees per acre in a stand if the ages and site qualities of the stand are given. Judged from this point of view the success of any curve may be measured in terms of the accuracy which results from such a use. If the curve is based on an adequate sample of the material to which it is to be applied, the accuracy of such application is essentially the same as that which will result from applying the curve to the basic data. The latter may be readily determined by estimating values for each datum and then comparing these estimates with the values actually measured. Only exceptionally in forestry has accuracy been thus determined. In most cases indeed the curve drawing has been so carried out as to obscure it completely.

To illustrate this statement, Figure 1, based on hypothetical measurements of age and diameter, has been prepared. The lower portion, C, represents the usual manner in which curves are drawn and presented. The individual measurements have been sorted into 10-year age classes and average diameters for these age classes have been plotted and assigned weights equal to the number of data in each. The advantage in this procedure is that the drawing of the curve is easier, particularly if the individual values are widely scattered. In Figure 1, B, the individual values have been plotted. A zone of points results which, in this instance, has such narrow limits that the curve could have been drawn readily without computing and plotting averages. Figure 1, A, shows a case of wider dispersion where the curve location would have been somewhat less certain, and where averaging is clearly advisable. The point of interest, however, is that the values used in A and B have been so chosen that C may be derived from either. If average points are plotted, the two cases will appear identical, yet obviously they are not. It is clear that in case B the diameter of any individual tree may be estimated by means of the curve with far less average error than in case A. Curve B must then be considered the more effective of the two. Its greater reliability is completely concealed when form C is used.

A measure of the accuracy of estimate, the average error, can readily be obtained for either A or B. Its computation is illustrated by Table 1. Columns 1, 2, and 6 show the age and diameter as actually measured. Columns 3 and 7 show the diameter as estimated from the age by means of the curve of Figure 1, A. In columns 4 and 8 are entered the residuals, a term which is used for the differences between the measured and estimated values. Where the measured exceeds the estimated value the residual is considered positive, and otherwise negative. The sum of columns 4 or 8 (disregarding signs) divided by the number of observations is the average error. Column 5 (and 9) should be disregarded for the time being.

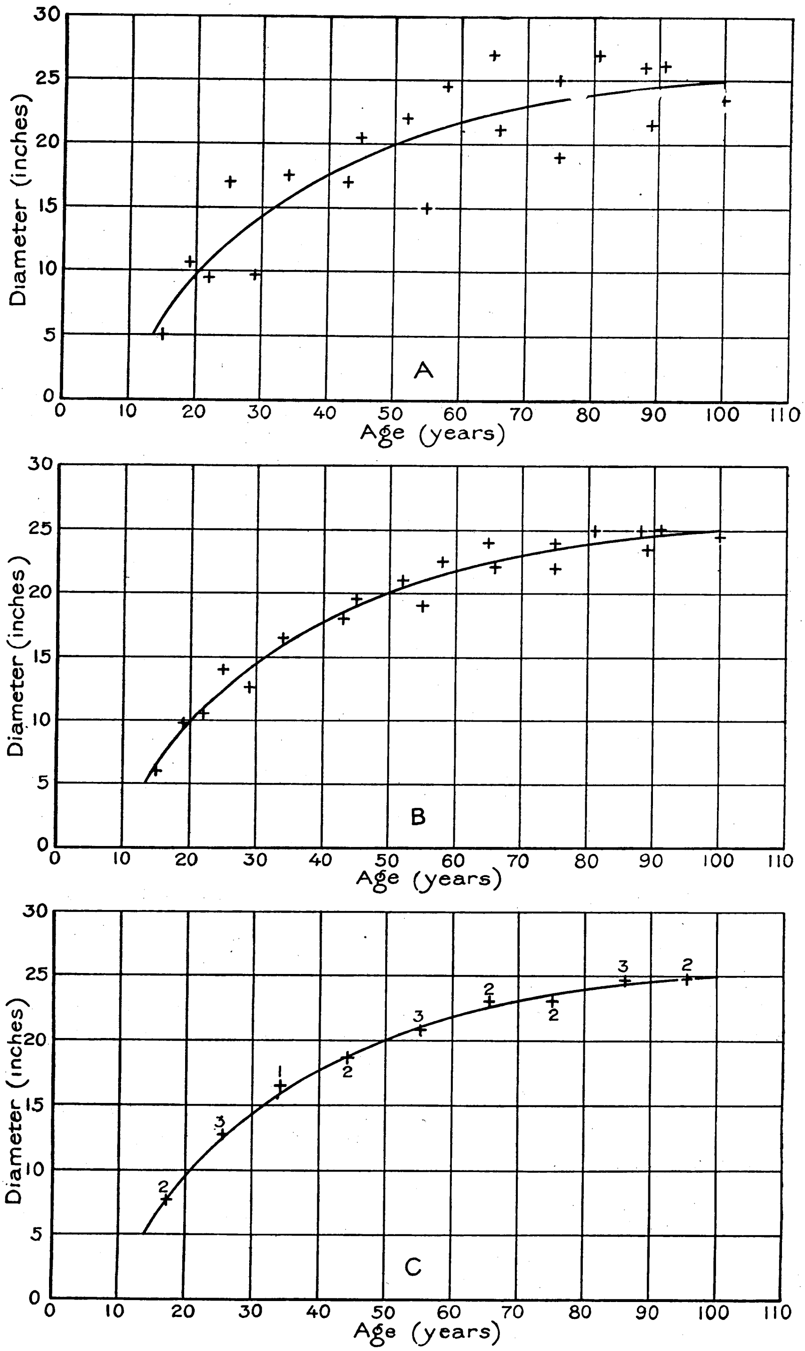

FIGURE 1.—The closeness with which a curve fits all its data may be concealed by the use of averages. Curve C may be derived from either A or B, yet B is obviously superior to A

TABLE 1.—*Computation of average error and standard error for Figure 1, A and B*

Age	Curve A				Curve B			
	Measured diameter	Diameter estimated from curve	Residual	Residual squared	Measured diameter	Diameter estimated from curve	Residual	Residual squared
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
15	5.0	6.5	−1.5	2.25	6.0	6.5	−0.5	0.25
19	10.7	9.2	+1.5	2.25	9.7	9.2	+.5	.25
22	9.5	11.0	−1.5	2.25	10.5	11.0	−.5	.25
25	17.0	12.5	+4.5	20.25	14.0	12.5	+1.5	2.25
29	9.7	14.2	−4.5	20.25	12.7	14.2	−1.5	2.25
34	17.5	16.0	+1.5	2.25	16.5	16.0	+.5	.25
43	17.0	18.5	−1.5	2.25	18.0	18.5	−.5	.25
45	20.5	19.0	+1.5	2.25	19.5	19.0	+.5	.25
52	22.0	20.5	+1.5	2.25	21.0	20.5	+.5	.25
55	15.0	21.0	−6.0	36.00	19.0	21.0	−2.0	4.00
58	24.5	21.5	+3.0	9.00	22.5	21.5	+1.0	1.00
65	27.0	22.5	+4.5	20.25	24.0	22.5	+1.5	2.25
66	21.1	22.6	−1.5	2.25	22.1	22.6	−.5	.25
75	19.0	23.5	−4.5	20.25	22.0	23.5	−1.5	2.25
75	25.0	23.5	+1.5	2.25	24.0	23.5	+.5	.25
81	27.0	24.0	+3.0	9.00	25.0	24.0	+1.0	1.00
88	26.0	24.5	+1.5	2.25	25.0	24.5	+.5	.25
89	21.5	24.5	−3.0	9.00	23.5	24.5	−1.0	1.00
91	26.1	24.6	+1.5	2.25	25.1	24.6	+.5	.25
100	23.5	25.0	−1.5	2.25	24.5	25.0	−.5	.25
Total	384.6	384.6	.0	171.00	384.6	384.6	.0	19.00
Average	19.2	19.2	.0	8.55	19.2	19.2	.0	.95
Total, disregarding signs			51.0				17.0	
Average, disregarding signs			2.55				.85	
Standard error				$\sqrt{8.55}$				$\sqrt{.95}$
Or				2.924				.975

It should be noted first that in both cases the total of the estimated diameters (columns 3 and 7, which are of course identical, since the curves are the same) equals the total of the measured diameters (columns 2 and 6). This is a fundamental criterion [4] of a correctly fitted curve. An inevitable corollary to this fact is that the sums of the positive and negative residuals are equal. Hence, the algebraic sums of columns 4 and 8 are zero. In actual practice this last fact is the more usable, and in curve fitting the differences between the curve and all points above it may be quickly summed and compared with a similar value for all points below the curve. If they are not approximately equal, the curve should be so shifted as to eliminate any material difference. This test can be applied equally well to a graph of form 1, C, by weighting (multiplying) each difference by the number of observations involved therein.

The dissimilarity between the data for A and B is brought out by the average residuals (last values in columns 4 and 8), disregarding signs in the computation. It is apparent that the average error of estimate is nearly three times as great in case A as in case B, for if curve A is used the diameters estimated will be in error by an average of 2.55 inches, while if B is used this average error will be reduced to 0.85 inch.

A somewhat better measure of accuracy than this has been devised. It is called the "standard error" [5] and its calculation is illustrated by

[4] Although a curve is thus accurately balanced it may still be so tilted or poorly shaped that it fails to fit the data.

[5] There is some difference in the usage of statistical terms. Standard error is sometimes used as synonyomous with standard deviation of a mean.

columns 5 and 9. In computing this each residual is squared, and these squares are totaled. Their sum is then divided by the number of values involved and the square root extracted. A formula which expresses this is—

$$SE=\sqrt{\frac{\text{Sum } e^2}{N}} \qquad \text{(I)}$$

where SE is the symbol for standard error, Sum signifies the sum of all the values for the expression immediately following, e is the residual and N is the number of observations used. The standard error is superior to the average error because it involves the generally accepted theory of least squares. On this theory a properly fitted curve of given form has a minimum standard error rather than a minimum average error. In most cases where the dispersion of points above the curve is similar to that below the curve, the standard error is approximately one and one-fourth times the average error, but it is often unsafe to calculate it in this way. In the present instance the standard errors thus calculated would be 3.19 for A, and 1.06 for B. These values differ appreciably from the more accurately determined standard errors for A and B, which are calculated by means of columns 5 and 9, Table 1, and Formula I—

$$\text{(A)} \qquad SE=\sqrt{\frac{171.00}{20}}=2.924$$

$$\text{(B)} \qquad SE=\sqrt{\frac{19.00}{20}}=0.975$$

Like the average error, the standard error of A is approximately three times that of B, so that in the present case the same relative result is obtained by either method of computation.

Either the standard error or the average error, then, measures the accuracy of a curve as a means of predicting values of one variable from values of another. It must not be assumed, however, that small standard errors necessarily mean that the curve which has been drawn is useful or that large standard errors imply that it is futile. Its effectiveness depends not only on the standard error but also on how much variation was originally present. If the variable investigated was relatively stable in its values, a small standard error might be obtained even with very bad curve fitting, or even by estimating without any curve at all by using the arithmetic average. The graphic equivalent of this average is a horizontal straight line through its value. To use this obviously disregards any possible variation associated with the independent variable.

What has been accomplished by a curve is, therefore, better judged by the relation between the scatter about the curve and the scatter of the data as a whole about its arithmetic mean. The measure of the scatter about the arithmetic mean is the standard error of the points about the horizontal straight line through the average. This special case of standard error is called the standard deviation.

STANDARD DEVIATION

For the example already cited, the calculation of the standard deviation is shown in Table 2. In the first and fourth columns the measured diameters are entered. These columns are summed and the averages obtained. The residuals (the differences between the individual values and the average), which are in this case called "deviations," are next computed. The calculation of the standard deviation then is exactly similar to that of the standard error, the formula being—

$$SD=\sqrt{\frac{\text{Sum } d^2}{N}} \qquad \text{(II)}$$

where *SD* is the symbol for standard deviation, *d* signifies deviation and *N*, number of observations. The resemblance to Formula I is obvious. The computations in the present case are in columns 2, 3, 5, and 6.

TABLE 2.—*Computation of standard deviations for data used in Figure 1, A and B*

	Curve A			Curve B		
	Measured diameter	Deviation from average	Deviation squared	Measured diameter	Deviation from average	Deviation squared
	(1)	(2)	(3)	(4)	(5)	(6)
	5.0	−14.2	201.64	6.0	−13.2	174.24
	10.7	−8.5	72.25	9.7	−9.5	90.25
	9.5	−9.7	94.09	10.5	−8.7	75.69
	17.0	−2.2	4.84	14.0	−5.2	27.04
	9.7	−9.5	90.25	12.7	−6.5	42.25
	17.5	−1.7	2.89	16.5	−2.7	7.29
	17.0	−2.2	4.84	18.0	−1.2	1.44
	20.5	+1.3	1.69	19.5	+.3	.09
	22.0	+2.8	7.84	21.0	+1.8	3.24
	15.0	−4.2	17.64	19.0	−.2	.04
	24.5	+5.3	28.09	22.5	+3.3	10.89
	27.0	+7.8	60.84	24.0	+4.8	23.04
	21.1	+1.9	3.61	22.1	+2.9	8.41
	19.0	− .2	.04	22.0	+2.8	7.84
	25.0	+5.8	33.64	24.0	+4.8	23.04
	27.0	+7.8	60.84	25.0	+5.8	33.64
	26.0	+6.8	46.24	25.0	+5.8	33.64
	21.5	+2.3	5.29	23.5	+4.3	18.49
	26.1	+6.9	47.61	25.1	+5.9	34.81
	23.5	+4.3	18.49	24.5	+5.3	28.09
Total	384.6	------------	802.66	384.6	------------	643.46
Average	19.2	------------	40.13	19.2	------------	32.17
Standard deviation		------------	$\sqrt{40.13}$	------------	------------	$\sqrt{32.17}$
Or		------------	6.33	------------	------------	5.67

Were nothing known about age, the only possible method of estimating the diameter of trees from the material at hand would be to guess that each one was of average diameter; in other words to estimate by means of a horizontal line such as has been mentioned. Individual errors would be large, although in the long run compensating, if the material is an adequate sample of the diameters. The use of the diameter–age curve permits a decided improvement[6] in estimating as is shown by the fact that the standard error is only 2.924 in case A, while the corresponding standard deviation is 6.33. Similar values in case B are 0.975 and 5.67.

[6] This estimate of improvement is conservative. It is relatively easy to secure sample measurements of diameter and age, which adequately represent their relation in the stand in which they are taken. Such measurements, however, may be entirely inadequate for determining either the average diameter or average age. A sample suitable for determining such averages can best be obtained by a strip survey.

ALIENATION INDEX

The estimating is not perfect, of course, even where the curve is used, as is shown by the existence of these residual standard errors of 2.924 and 0.975. Estimates are, however, materially improved by using the curve, and it is helpful to measure this improvement. The ratio between the residual variation and the original variation is known as the "alienation index." Expressed as a formula—

$$AI=\frac{SE}{SD} \tag{III}$$

where AI is the symbol for alienation index, SE signifies standard error, and SD signifies standard deviation. It is obvious that the values of this index may range from 0 to 1.00.[7] In the present instances these alienation indices are (A) $\frac{2.924}{6.333}=0.462$, and (B) $\frac{0.975}{5.67}=$ 0.172. This means that 46.2 and 17.2 per cent of the variability in diameter is associated with factors other than age. Among these factors, however, errors in fitting the curve may be included.

CORRELATION INDEX

The "correlation index" is another and more commonly used measure of this improvement. It may be derived from the alienation index by means of the formula:

$$CI=\sqrt{1-(AI)^2} \tag{IV}$$

where CI signifies correlation index, and AI alienation index. If the alienation index is 1 (the maximum possible value) the correlation index is evidently 0, while if the alienation index is 0 (only possible where the standard error is 0), then the correlation index is 1. This fixes the limits of possible values for the correlation index. The alienation index measures the association between the dependent variable and other unconsidered factors; the correlation index only indicates the association between the dependent variable and that independent variable which was considered. Complete association is shown by a correlation index of 1, while entire absence of association is shown by a correlation index of 0. Intermediate values show partial association. Unfortunately, a correlation index of 0.50 does not mean that half of the variation present in one variable is associated with the other variable used. The best way to interpret a correlation index is to compute the corresponding alienation index. By substituting in the foregoing formula the values of the alienation indices, which are 0.462 and 0.172, in the present case, the correlation indices are found to be 0.887 and 0.985, respectively. To avoid computation, approximate values [8] may be read from Figure 2.

It will be seen that both the standard error and the alienation or correlation index are useful as measures of accomplishment and that neither gives complete information without the other. The standard

[7] A value greater than 1.00 can conceivably be obtained, but only through grotesque misfitting of a curve. Such a value indicates that the curve not only has no utility but is completely misleading and should be discarded.

[8] See also Miner (*30*).

error shows in absolute units how accurately the value of a variable may be estimated from values of another; it does not show whether an estimate might have been made as accurately or nearly as accurately without it. The alienation index is a more abstract value measuring the relative improvement of estimate consequent on the use of a certain independent variable, but it does not show the amount of error remaining. Both measures should ordinarily be calculated.

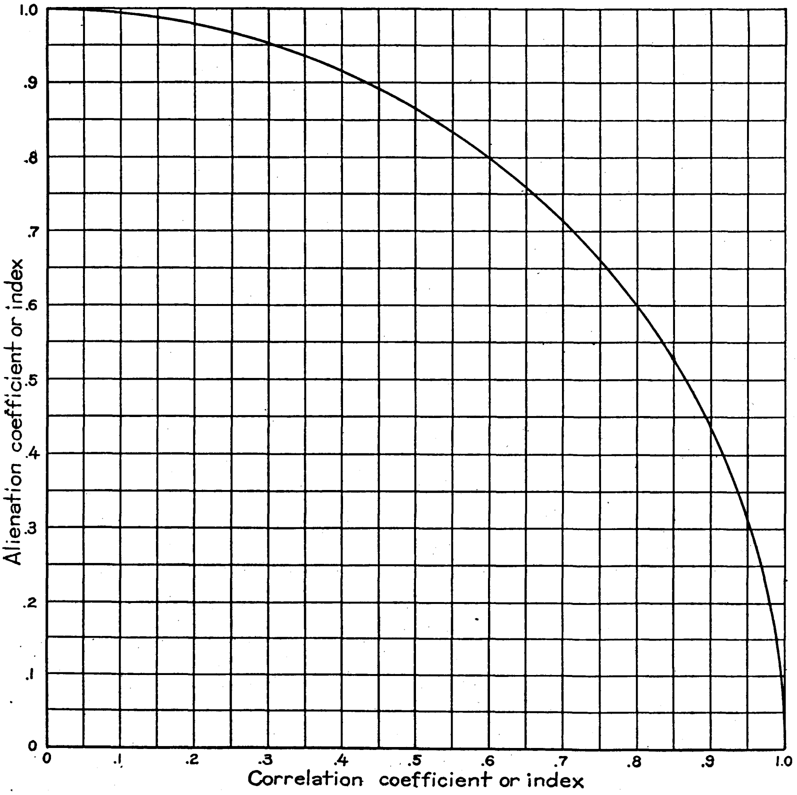

FIGURE 2.—The relation between alienation and correlation indices is shown by this curve. It may be used to obtain, without computation, approximate values of either when the other is known

MULTIPLE CORRELATION; IMPROVING THE ESTIMATE BY THE USE OF ADDITIONAL VARIABLES

Where the coefficient of alienation is high, the possibility of further improving the estimate by the use of one or more additional variables suggests itself. Volume tables afford a common example. These may be prepared on the basis of diameter alone, but the resulting standard error and alienation index will prove to be large. It is customary to include height as an additional independent variable. The graphic result of this is the familiar set of harmonized curves. The accuracy of estimate should be, and is, improved by this treatment. The calculation of the standard error, alienation index, and correlation index is entirely analogous to that of the 2-variable case just discussed. In reading the "estimated volumes from curve" in this case, it will be necessary to interpolate between the curves. This adds considerably to the labor involved.

For this and other reasons the graphic method is difficult where there are three variables and hardly suitable to problems involving more than three, although some effort has been made to adapt it to four, as for taper curves.

The terms "multiple alienation index" and "multiple correlation index" are used to distinguish such three or more variable cases from those where but two variables are involved.

It has now been shown that such statistical conceptions as the standard error and the alienation index may be applied to the curves or systems of harmonized curves with which foresters are familiar. It has also been shown that they give us a quantitative measure of the utility of these curves. Certain substitutes for free-hand curve drawing will next be described and the relative adequacy of the results discussed.

REGRESSION EQUATION

When data are plotted as in Figure 1, it may be that a straight line is defined instead of a curve. When this is so, statistical methods offer a purely mathematical means of locating, with rigorous accuracy, the best position of this straight line by calculating its equation known as the regression equation. Being the equation of a straight line it must be of the type—

$$Y=AX+B$$

where Y is the dependent variable, X the independent variable, and A and B are constants which must be determined from the data at hand.

These constants may be computed by the standard method of least squares, but the same results may be obtained by the application of the relatively simple product moments formula—

$$Y=M_Y+\frac{\text{Sum } d_X d_Y}{\text{Sum } d^2_X}(X-M_X) \qquad \text{(V)}$$

where X is the independent variable, Y the dependent variable, M_X the arithmetic mean of X, M_Y the arithmetic mean of Y, and d_X and d_Y the deviations of X and Y from their means.

In statistical work it is not uncommon to calculate this equation even where it is not definitely known that the relation is strictly rectilinear. This may be because any curvilinear trend present is poorly defined, or because the straight-line value is useful as a first approximation, as will later be explained. It is, therefore, not unreasonable to use the material already presented in Table 1 as an example of this process. The data for curve B only will be used.

The form of computation is shown in Table 3. Columns 1 and 4 list the measurements. Column 2 contains the deviations of the individual ages from their mean, with the signs noted. Column 3 contains these values squared. These two columns are similar to columns 5 and 6 in Table 2. In column 5 are listed similar deviations for diameter. Column 6 is not used in the present computation. In column 7 are the products, for each item, of its deviation in age and in diameter. As indicated at the bottom of the table, the equation becomes:

$$\text{Diameter} = 0.202\ \text{Age} + 7.85$$

Figure 3 shows the line representing this equation plotted through the zone of points upon which it is based.

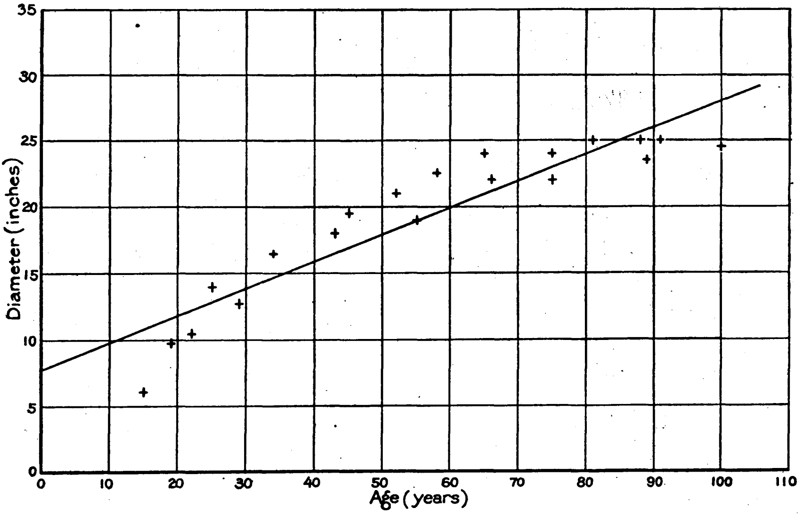

FIGURE 3.—The regression line for the data used in Figure 1, B. This is the straight line which best fits the plotted points, but it is obviously less satisfactory than the curve of Figure 1, B

TABLE 3.—*Calculation of regression equation from data of Figure 1, B*

	Age	Deviation of age from mean	Age deviation squared	Diameter	Deviation of diameter from mean	Diameter deviation squared	Product of diameter deviation and age deviation
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	15	−41.35	1,709.8	6.0	−13.2	174.24	+545.82
	19	−37.35	1,395.0	9.7	−9.5	90.25	+354.82
	22	−34.35	1,179.9	10.5	−8.7	75.69	+298.84
	25	−31.35	982.8	14.0	−5.2	27.04	+163.02
	29	−27.35	748.0	12.7	−6.5	42.25	+177.78
	34	−22.35	499.5	16.5	−2.7	7.29	+60.34
	43	−13.35	178.2	18.0	−1.2	1.44	+16.02
	45	−11.35	128.8	19.5	+.3	.09	−3.40
	52	−4.35	18.9	21.0	+1.8	3.24	−7.83
	55	−1.35	1.8	19.0	−.2	.04	+.27
	58	+1.65	2.7	22.5	+3.3	10.89	+5.44
	65	+8.65	74.8	24.0	+4.8	23.04	+41.52
	66	+9.65	93.1	22.1	+2.9	8.41	+27.98
	75	+18.65	347.8	22.0	+2.8	7.84	+52.22
	75	+18.65	347.8	24.0	+4.8	23.04	+89.52
	81	+24.65	607.6	25.0	+5.8	33.64	+142.97
	88	+31.65	1,001.7	25.0	+5.8	33.64	+183.57
	89	+32.65	1,066.0	23.5	+4.3	18.49	+140.40
	91	+34.65	1,200.6	25.1	+5.9	34.81	+204.44
	100	+43.65	1,905.3	24.5	+5.3	28.09	+231.34
Total	1,127	----------	13,490.1	384.6	----------	643.46	2,725.08
Mean	56.35	----------	----------	19.23	----------	----------	----------
Standard deviation	----------	----------	26.0	----------	----------	5.67	----------

$$\text{(V) Diameter} = M_{\text{Dia.}} + \frac{\text{Sum } d_{\text{Dia.}} . d_{\text{Age}}}{\text{Sum } d^2_{\text{Age}}} (\text{Age} - M_{\text{Age}})$$

$$= 19.23 + \frac{2{,}725.08}{13{,}490.1} (\text{Age} - 56.35)$$

$$= 0.202\ \text{Age} + 7.85$$

ALIENATION AND CORRELATION COEFFICIENTS

It is obvious that the curve in Figure 3 fits the points less well than did the curve of Figure 1. How much less can be determined quantitatively by comparing its standard error with that of the curve. This standard error is computed from estimated values read from the straight line or calculated directly from the equation.. The result, shown in Table 4, is 2.15. By equation III, then, the alienation index is $\frac{2.15}{5.67}=0.379$. By equation IV the corresponding correlation index is—

$$\sqrt{1-(0.379)^2}=0.925$$

In such a case as this, however, the alienation index and the correlation index are called the alienation coefficient and correlation coefficient. The only difference between the index and the coefficient is that the coefficient is based on a straight line instead of a curve. The coefficients may, therefore, be considered merely as special cases of the more general indices. They are, however, more widely used, largely because they can be computed directly by a shorter method which does not involve the use of estimated values of the independent variables. This method does not apply except where straight lines are used.

TABLE 4.—*Calculation of standard error from the regression equation, Diameter= 0.202 Age+7.85*

	Age	Measured diameter	Diameter estimated from regression equation	Residual	Residual squared
	(1)	(2)	(3)	(4)	(5)
	15	6.0	10.9	−4.9	24.01
	19	9.7	11.7	−2.0	4.00
	22	10.5	12.3	−1.8	3.24
	25	14.0	12.9	+1.1	1.21
	29	12.7	13.7	−1.0	1.00
	34	16.5	14.7	+1.8	3.24
	43	18.0	16.5	+1.5	2.25
	45	19.5	16.9	+2.6	6.76
	52	21.0	18.4	+2.6	6.76
	55	19.0	19.0	.0	.00
	58	22.5	19.6	+2.9	8.41
	65	24.0	21.0	+3.0	9.00
	66	22.1	21.2	+.9	.81
	75	22.0	23.0	−1.0	1.00
	75	24.0	23.0	+1.0	1.00
	81	25.0	24.2	+.8	.66
	88	25.0	25.6	−.6	.34
	89	23.5	25.8	−2.3	5.29
	91	25.1	26.2	−1.1	1.21
	100	24.5	28.0	−3.5	12.25
Total		384.6	384.6	.0	92.44
Average		19.23	19.23		4.62

Standard error=$\sqrt{4.62}$=2.15

The formula involved is—

$$AC_{XY}=\sqrt{1-\frac{\text{Sum}^2\ d_Xd_Y}{(\text{Sum}\ d^2{}_X)\ (\text{Sum}\ d^2{}_Y)}} \qquad \text{(VI)}$$

where AC_{XY} is the alienation coefficient between X and Y and the other symbols are as before. The values to be entered in this formual have already been computed in Table 3, to which column 6 has been added for this purpose. Using these values we have:

$$AC_{\text{Dia., Age}}=\sqrt{1-\frac{(2725.08)^2}{13490.1\times 643.46}}=0.380$$

This is approximately the same as the value found in the preceding paragraph (0.379). The small difference is due merely to failure to carry out the computations to enough significant figures.

It is interesting to compare the alienation coefficient based on the regression straight line with the alienation index based on the curve of Figure 1, B. On theoretical grounds it is obvious that since the curve is the better estimating mechanism its alienation index should be the lower. This is found to be true, its value being 0.172 as compared with that of 0.379 for the alienation coefficient. It may then be said that the curve is slightly over twice as effective as the line.

For this reason, the correlation coefficient (*CC*), which is based on the straight line, will naturally be lower than the correlation index. Its value is 0.925, as compared with 0.985. Obviously the correlation values convey a less definite sense of the relative efficiency of line and curve than do the alienation values. It should be noted that the alienation coefficient (and the correlation coefficient associated therewith) is always a conservative estimate of the importance of one variable in determining values of another.

It is customary to use a plus or minus sign before the correlation coefficient to indicate the direction of slope of the regression line. The plus sign implies a rising line, or in other words, that increases in one variable are associated with increases in the other, while a minus sign implies the reverse. If Formula VI is used, the sign of the term "Sum d_Xd_Y" (before squaring), indicates the sign of the coefficient. It is not customary to attach a sign to the correlation index since it is both positive and negative when part of the curve rises and part falls.

RELATIVE ADVANTAGES AND DISADVANTAGES OF THE STATISTICAL AND THE GRAPHIC METHODS

Although the nongraphic methods, considered alone, are more laborious than simple curve drawing, they are actually less laborious if to the work of drawing the curves is added that of checking their adequacy. The statistical methods almost simultaneously produce both a predicting mechanism and a criterion of its effectiveness. Furthermore the result obtained is free from personal bias, and different workers using the same data will arrive at identical conclusions. On the other hand, the greater flexibility of the graphic method will permit greater accuracy in the vast majority of instances where only two variables are involved. But where three or more variables must be considered, the purely graphic method is less successful; its advantages of flexibility and accuracy are less marked, and its difficulties are intensified. It follows that a combination of the two is desirable in such cases.

GRAPHIC METHODS OF CHECKING CURVE FIT

RESIDUAL CURVE

Before passing on to a more detailed consideration of multiple linear correlation, certain processes will be described which have little or no practical utility in connection with 2-variable problems, but which are more readily explained in connection with them. These preliminary considerations will be of great service in facilitating the discussion of multiple correlation problems.

In Table 1, column 8, have been listed the residuals of curve B. Where a point is above the curve a plus sign has been used, and where it is below the curve a minus. These residuals may be plotted as a dependent variable against the original independent variable, which in this case is age. The result is shown in Figure 4, A. It will be seen that although the plotted points of the original data defined a curve (Fig. 1, B), the plotted deviations from this curve define a horizontal straight line. (Fig. 4, A.) It should be clear without detailed proof that this always results from correct curve fitting, and it follows that this graph of residuals may be used as a means of checking the correctness of any curve. Another way of stating the same thing is that the residuals should not be correlated with the independent variable. A test of this might equally well be made mathematically by the method outlined on page 10. If this were done, the evidence of perfect curve fitting would be an alienation coefficient of 1.00 and a regression equation which represented a horizontal line. While sound, neither of these tests is of sufficient value to justify the labor involved.

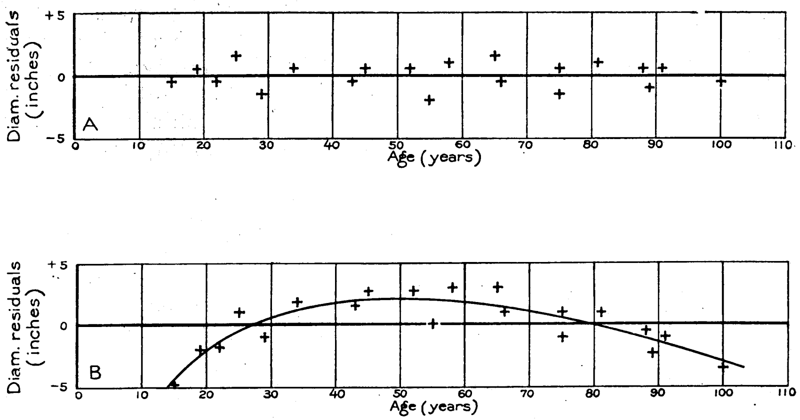

FIGURE 4.—The residual curves corresponding to Figure 1, B, and Figure 3. Correctly fitted curves should result in horizontal straight lines, signifying no correlation between residuals and independent variables

As a case of poor curve fitting to compare with the above, Figure 3 may be considered. The corresponding residuals are given in Table 4, column 4. If these are plotted Figure 4, B, results and a well defined curve appears. If this curve is drawn and values read from it are added to values read from Figure 3, the sums, when plotted, will produce the curve of Figure 1, B. This illustrates an indirect method of curve fitting in which the regression line is, as it were, used as a first approximation to the truth, and is afterward converted into the true curve by means of the residuals. In the present instance this method is of no value, producing no better results than the direct method, and considerably lengthening the work. In problems involving several variables, however, a slight modification of the principle involved will be of great utility.

The alienation coefficient of the residuals in Figure 3 will be found to be 1, and their regression line will be horizontal in spite of the obviously poor fit. This indicates merely that the straight line of this figure is the best fitting straight line, a fact which does not preclude the possibility of improving on the straight line by adding curvature. This illustrates the limitations of this mathematical treatment. It will not distinguish at all between such cases as those illustrated in Figures 4, A and B. In fact it will show errors only in the tilt of the line or curve.

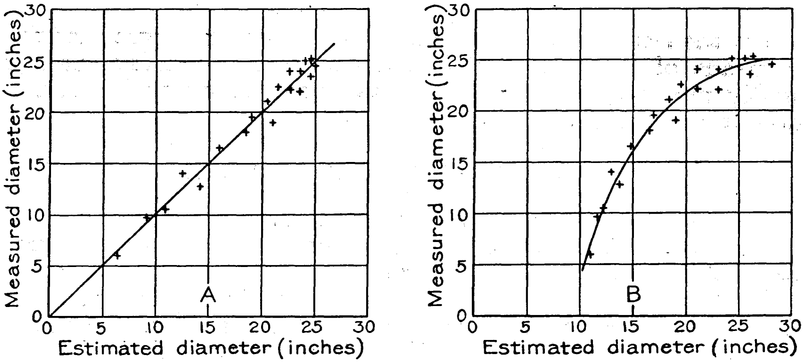

FIGURE 5.—Curves showing comparison between measured and estimated values for correctly and incorrectly fitted curves. A is derived from Figure 1, B, and B from Figure 3. Correct curves result in 45 degree straight lines passing through the origin

CURVE OF RELATION BETWEEN MEASURED AND ESTIMATED VALUES

The second type of derived curve, similar in purpose and usefulness, is prepared by plotting measured against estimated values of the variable, using the same scale for each axis. If this be done for the same data as before, i. e., that of Figure 1, B, and Figure 3 (the values used being listed in columns 6 and 7, Table 1, and columns 2 and 3, Table 4), A and B of Figure 5 result. It will be seen that a correctly fitted curve by this treatment results in a 45-degree straight line radiating from the origin as in Figure 5, A. Any other line, or a curve such as Figure 5, B, indicates erroneous fitting. The interpretation of various possibilities is as follows:

A. A 45-degree line not radiating from the origin means that the fitted curve should be raised or lowered without changing its tilt or form.

B. A straight line not 45° indicates that the tilt should be changed without changing the form.

C. A curve indicates the desirability of a change in form.

As in the previous case, Figure 5, B, may be used in conjunction with the straight line on which it is based (fig. 3) to produce the true curve. The process of adjustment, however, differs and is illustrated in Table 5. In this the column 1 values are arbitrarily selected. Column 2 contains values read from Figure 3 (the straight line) corresponding to the column 1 values, while column 3 contains values read from the curve of Figure 5, B, corresponding to the column 2 values. The result is practically identical with the curve of Figure 1, B.

TABLE 5.—*Correction of regression line by means of Figure 5, B, to produce curves similar to Figure 1*

Age	Estimated diameter [1]	Corresponding measured diameter [2]	Age	Estimated diameter [1]	Corresponding measured diameter [2]	Age	Estimated diameter [1]	Corresponding measured diameter [2]
Years	*Inches*	*Inches*	*Years*	*Inches*	*Inches*	*Years*	*Inches*	*Inches*
20	11.8	9.6	50	17.9	20.0	80	24.0	24.0
30	13.9	14.3	60	20.0	21.7	90	26.1	24.7
40	15.9	17.6	70	22.0	23.1	100	28.1	24.9

[1] From Figure 3. [2] From Figure 5, B.

Like the last process, this method of preparing curves indirectly by means of the regression straight line has no utility here, but slightly modified it also will be used with advantage in multiple-correlation problems.

MULTIPLE REGRESSION EQUATION

Where more than three variables are involved in a problem, a nongraphic method is available which is closely parallel to that already described for two variables. Since a form analogous to a straight line is assumed, the equation employed remains in the first degree, but must provide for additional independent variables. Its form is (for three variables) $W=AX+BY+K$; (for four variables) $W=AX+BY+CZ+K$, etc., where W is the dependent variable, X, Y, and Z are the independent variables, and A, B, C, and K are constants to be determined from the available data.

The determining of these constants is best done by the method of least squares. It may be accomplished through the use of the following procedure.

It is first necessary to rewrite the equation in slightly different symbols. This revised form is quite closely analogous to that on page 10 for two variables, thus—

$$W=M_W+B_{WX}\frac{SD_W}{SD_X}(X-M_X)+B_{WY}\frac{SD_W}{SD_Y}(Y-M_Y) \qquad \text{(VII)}$$

where W is the dependent variable; M_W, M_X, and M_Y are the mean values of W, X, and Y; SD_W, SD_X, and SD_Y are the standard deviations of W, X, and Y; and B_{WX} and B_{WY} are coefficients to be determined. The means and standard deviations are determind by the method already described (p. 6), but the determination of the coefficients[9] B_{WX}, etc., requires the use of two or more "normal

[9] CC_{WX} and CC_{XW} are identical, but B_{WX} and B_{XW} are not. The correct signs of the correlation coefficients must be used (see p. 13).

equations," which are as follows, VIII–A serving for three variables, and VIII–B for four variables—

$$\left.\begin{array}{rcl} B_{WX}+CC_{XY}B_{WY} & = & CC_{WX} \\ CC_{YX}B_{WX}+\qquad\quad B_{WY} & = & CC_{WY}\end{array}\right\}\text{(VIII–A)}$$

$$\left.\begin{array}{rcl} B_{WX}+CC_{XY}B_{WY}+CC_{XZ}B_{WZ} & = & CC_{WX} \\ CC_{YX}B_{WX}+\qquad\quad B_{WY}+CC_{YZ}B_{WZ} & = & CC_{WY} \\ CC_{ZX}B_{WX}+CC_{ZY}B_{WY}+\qquad\quad B_{WZ} & = & CC_{WZ}\end{array}\right\}\text{(VIII–B)}$$

etc., etc., each additional variable adding one equation and lengthening each equation by one term. By observing the sequence of terms it is easy to write the four equations for a problem with five variables, etc.

To illustrate this method it is necessary to select a problem involving more than two variables, such as that of the relation between bark thickness, d. b. h. (diameter breast high), and height of second-growth longleaf pine trees. A considerable number of data will be used (564 tree measurements) and it would be out of the question to list them. The computations of averages, standard deviations, and alienation coefficients involve no new principle. These values, therefore, may be taken as a starting point. They are:

	Mean	Standard deviation	
Bark thickness (W)	0. 605	0. 185	
Diameter (X)	7. 01	2. 67	
Height (Y)	54. 99	16. 7	
Alienation coefficients, bark and diameter (AC_{WX})			0. 893
Alienation coefficients, bark and height (AC_{WY})			. 968
Alienation coefficients, diameter and height (AC_{XY})			. 524

From these the following correlation coefficients were calculated:

Bark and diameter (CC_{WX})	+0. 450
Bark and height (CC_{WY})	+. 251
Diameter and height (CC_{XY})	+. 852

Substituting these values in the normal equations (VIII–A) we have—

$$B_{WX}+0.852\ B_{WY}=0.450$$

$$0.852\ B_{WX}+B_{WY}=0.251$$

These equations must now be solved simultaneously to obtain the values of B_{WX} and B_{WY}. For example, we may multiply the second by 0.852 and then subtract it from the first, thus—

$$\begin{array}{rl} B_{WX}+0.852\ B_{WY} & =0.450 \\ \underline{0.726\ B_{WX}+0.852\ B_{WY}} & \underline{=0.214} \\ 0.274\ B_{WX}\qquad\qquad & =0.236 \\ B_{WX}\qquad\qquad & =0.861 \end{array}$$

Then, if this value of B_{WX} is substituted in the first equation, it becomes—

$$\begin{array}{rl} 0.861+0.852\ B_{WY} & =\ 0.450 \\ 0.852\ B_{WY} & =-.411 \\ B_{WY} & =-.482 \end{array}$$

The values of the coefficients may now be substituted in Equation VII, as may also the given values for means and standard deviations. This will give—

$$W=0.605+0.861\,\frac{0.185}{2.67}\,(X-7.01)-0.482\,\frac{0.185}{16.7}\,(Y-54.99)$$
$$=0.0596X-0.00534Y+0.481$$

which is the required multiple regression equation.

The accuracy of this equation as a mechanism for predicting bark thickness can be measured, as in the cases previously illustrated, by computing estimated bark thickness for each of the 564 trees measured and then, by comparing estimated and measured values, calculating the standard error. From this in turn the multiple coefficient of alienation and the multiple coefficient of correlation can be derived. In the present instance the standard error is 0.158 inches. The multiple alienation coefficient is then—

$$\frac{SE}{SD}=\frac{0.158}{0.185}=0.854,$$

and the multiple correlation coefficient is—

$$\sqrt{1-(0.854)^2}=0.520$$

This method brings out best the principles involved, but time can be saved by calculating the alienation coefficient directly and then working backwards to the other values needed. The equation which may be used is—

$$AC_{W(XYZ--)}=\sqrt{1-(B_{WX}CC_{WX}+B_{WY}CC_{WY}+B_{WZ}CC_{WZ}+---)}\quad\text{(IX)}$$

where $AC_{W(XYZ--)}$ is the multiple alienation coefficient between W and X, Y, Z, etc. In the present instance this becomes—

$$AC_{W(XY)}=\sqrt{1-[(0.861\times0.450)+(-0.482\times0.251)]}=0.856$$

The corresponding correlation coefficient is $\sqrt{1-(0.856)^2}=0.517$. Since the standard deviation of the bark thickness is 0.185, the standard error may be obtained by means of Equation III—

$$AC=\frac{SE}{SD}$$

$$0.856=\frac{SE}{0.185}$$

$$SE=0.158$$

It will be seen that the values are approximately the same as before but that the computation is simple and brief, an estimate of each individual bark thickness being unnecessary. The small difference in the alienation coefficient could be eliminated by retaining more significant figures.

The alienation coefficient here is high. Of the variation in bark thickness 85.6 per cent is associated with factors other than diameter and height. As a predicting medium the regression equation obviously leaves much to be desired. It is not clear, however, how much of this high value is due to other unevaluated variables and how much is the result of assuming that the equation is linear. It has been seen that the alienation index may be much lower than the alienation coefficient where proper curves are used. This situation then emphasizes the need for improved methods of working with multiple curvilinear correlation, and it is such a new method that it is the primary purpose of this bulletin to present.

The linear regression equation shows that both height and diameter have some influence on bark thickness and permits an approximate appraisal of their relative importance. Moreover, this information can not readily be obtained by simple graphic methods. To illustrate this, Figure 6 is presented to show the measurements involved in this study plotted in the customary way over both diameter and height. While the points indicate some curvilinearity, in neither case can a straight line be considered a poor interpretation. One would conclude from examining these graphs that increases in height and diameter both are associated with an increase in bark thickness.

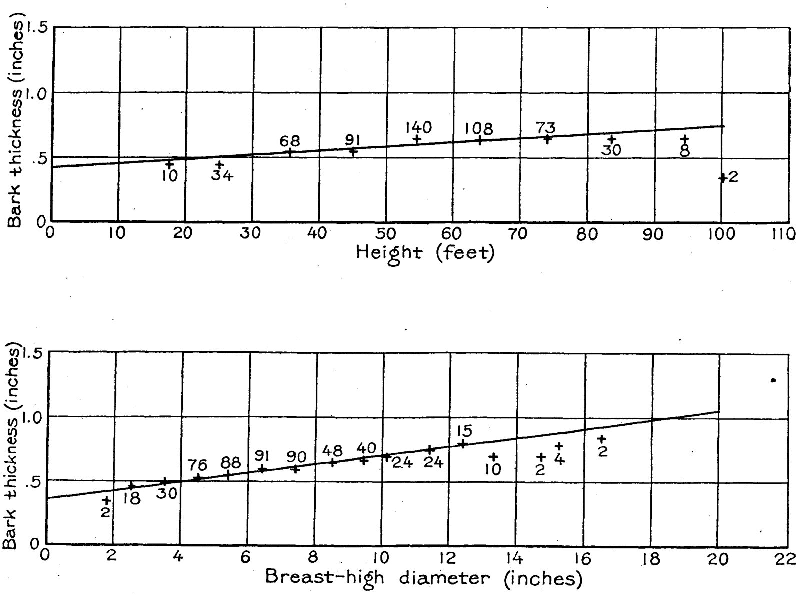

FIGURE 6.—Relation of bark thickness to height and bark thickness to diameter as brought out by 2-variable graphs. The correlation in each case is apparently positive. Compare with Figure 7

However, the multiple regression equation (p. 18)—

$$\text{Bark thickness} = 0.0596 \text{ diameter} - 0.00534 \text{ height} + 0.481$$

shows that while an increase in diameter is associated with an increase in bark, the larger the value used for height, the smaller the value which will result for bark thickness. This is an apparent contradiction to the conclusions drawn from Figure 6, B. To understand how these conclusions may be reconciled it is necessary to plot this equation, as in Figure 7, which was prepared by substituting for diameter the arbitrary values 2, 4, 6, etc., and then plotting the resulting series of straight lines. It will be seen that each individual line falls as the height increases, but that the series of lines viewed as a whole tends to rise. In other words, the net effect of increase in height, when diameter does not vary, is to decrease bark thickness; but when diameter is allowed to increase with height, as it normally does, the trees will have thicker bark, not because of their height but because of their diameter. Clearly both graphs represent the truth, but Figure 7 is a far more complete statement and is, in addition, a better means of predicting bark thickness. This same information can be obtained directly from the regression equation without plotting the graph. The plus sign before the coefficient of diameter (+0.0596) means that the net effect of diameter, i. e., the effect of diameter when height is constant, is to increase bark thickness, while the minus sign before the coefficient of height (−0.00534), means that the net effect of height is to decrease bark thickness. Furthermore, since the coefficient of diameter is the larger, diameter is a more important factor than height. In comparing the coefficients in such cases, however, the range of values of the variables must be borne in mind. If diameter, for example, ranges from 2 to 16 inches, the effect of diameter will cause bark thickness to vary from 0.0596×2=0.12 to 0.0596×16=0.95, a difference of 0.83 inch. On the other hand, if height ranges from 20 to 100 feet, its effect will range from 0.00534×20=0.11 to 0.00534×100=0.53, a difference of 0.42 inch. The maximum difference due to diameter is, therefore, only about twice as great as that due to height.

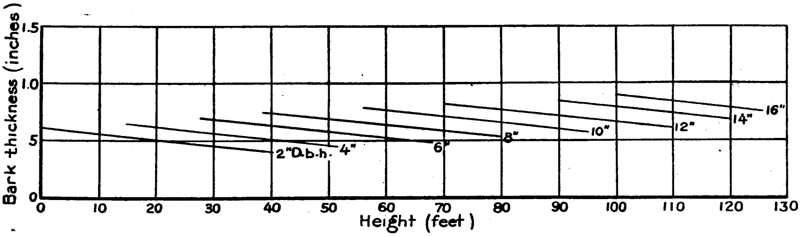

FIGURE 7.—Relation of bark thickness to height and diameter as brought out by the multiple regression equation (linear). Correlation between bark thickness and height is now seen to be negative. Compare with Figure 6

Another even better method of analyzing these relations is to compare the alienation coefficients and standard errors of the three regression equations which are based on height alone, on diameter alone, and on both. On pages 17 and 18 the corresponding alienation coefficients have been given; from these, standard errors may be computed in the usual way, as, for example:

Variables	Alienation coefficient	Standard error
Bark and height	0. 968	0. 179
Bark and diameter	. 893	. 165
Bark, diameter, and height	. 856	. 158

It will be seen that while all these alienation coefficients and standard errors are high the use of diameter alone reduces them considerably below what is obtainable by the use of height alone, and the use of both factors is marked by a still further improvement of slightly less magnitude. The standard error is improved by a reduction of 0.014 inch through using diameter instead of height, and an additional 0.007 by using both. Information of this sort should make possible an intelligent decision in any given problem as to what factors should be used, or whether a further search, either for additional variables or for curvilinearity in the relations may be desirable.

CURVILINEAR MULTIPLE CORRELATION

It should by now be obvious that a method of handling curvilinear multiple correlation is needed in many cases. As has already been said, where no more than three variables are involved, harmonized curves drawn by the conventional method long employed by foresters offer a solution. In the case just described, for example, much the same conclusions might have been reached by this method. In practice, however, this method is far from satisfactory because, (1) a very large number of data are required for satisfactory curves, and, (2) it is next to impossible to keep track adequately of weights during the construction of the second and subsequent sets of curves. As a result, in actual practice this method is becoming discredited even for 3-variable problems. As has already been stated, it is unusable for problems involving four or more variables.

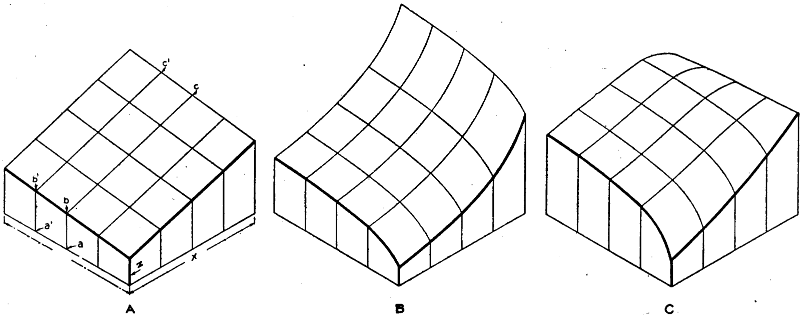

FIGURE 8.—Various types of correlation surfaces. A is a surface of the type assumed in ordinary linear multiple correlation. The equation is of the type $Z=AX+BY+C$. B is a correlation surface where the regression equation is of the type $Z=f_1(X)+f_2(Y)+C$. C is a surface where the regression equation is of the type $f_0(Z)=f_1(X)+f_2(Y)+C$

Clearly then, a mathematical method is needed. The chief difficulty is that the type of equation involved is usually unknown. In instances where it may be predicted, the least-squares method is available; but this is rarely the case in forestry problems.

Figure 8 illustrates three geometrically different types of relationship where three variables are involved. Just as a 2-variable equation may be considered as geometrically equivalent to a line, so a 3-variable equation may be considered as geometrically equivalent to a surface. If a 2-variable equation of the first degree may be represented by a straight line, a 3-variable equation of the first degree may be represented by a plane surface. Figure 8, A represents such a surface. The independent variables are assigned values on the two horizontal axes, and the dependent variable is measured vertically. Its equation is—

$$Z=AX+BY+C$$

Figures 8, B and C, represent two cases of nonlinear equations. Figure 8, B, represents an equation of the form—

$$Z = f_1(X) + f_2(Y) + C$$

where $f_1(\ )$ and $f_2(\ )$ signify "any function of."

The substitution, in such an equation, of a series of values for X (or Y), is equivalent to intersecting the surface by a series of parallel vertical planes. The lines of intersection in Figure 8, A, are *bc*, *b'c'*, etc., and these are analogous to the harmonized curves of the conventional graphic method. It is obvious, therefore, that in the present instance these harmonized curves are a series of parallel straight lines, and this is true whether the intersecting planes be parallel to the X or to the Y axis. In Figure 8, B, however, the resulting harmonized curves are nonlinear but still "parallel." The curve systems with which foresters are accustomed to deal are seldom of this type, and it, therefore, follows that a more flexible type of equation must be used. If it is assumed, however, that—

$$f_0(Z) = f_1(X) + f_2(Y) + C$$

the geometrical analogy may appear like Figure 8, C, and in this the harmonized curves are no longer necessarily parallel, because of the functional character of the dependent variable.

The appropriate method for such cases can perhaps best be explained by a concrete example. To permit a comparison of the relations empirically obtained with the true relations a hypothetical case will be set up by means of the assumed equation—

$$\left(\frac{Z}{4}\right)^2 = \sqrt{X} + 10 \log Y$$

It will be seen that this has been so chosen as to conform to the more general type just defined. Table 6 shows 30 sets of observations, the values of X and Y being selected at random, and the corresponding value of Z being calculated (to the nearest unit) by means of this equation. (These calculated values are to be considered equivalent to the measured values of an actual problem based on field data.) The problem is to find the essential equivalent of this equation by means of these 30 sets of values. With so few data it would be futile to attempt it by the conventional graphic method of harmonized curves. The method of attack will be, first, to determine the multiple regression equation and then later, by a series of successive approximations based on an analysis of the residuals, to modify the graphic equivalent of the equation by introducing whatever curvilinearity is present. The procedure will be described step by step.

TABLE 6.—*Data of example based on equation,* $\left(\frac{Z}{4}\right)^2=\sqrt{X}+10\ \log Y$, *and successive estimates of Z obtained by the multiple curvilinear correlation method*

Assumed values		Calculated values	Estimated values of *Z*, from alinement charts								Residuals								
X	*Y*	*Z*	First estimate	Second estimate	Third estimate	Fourth estimate	Fifth estimate	Sixth estimate	Seventh estimate	Eighth estimate	First estimate	Second estimate	Third estimate	Fourth estimate	Fifth estmate	Sixth estimate	Seventh estimate		Eighth estimate
																	A	B	
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)		(19)
11	10	15	13.8	15.0	15.0	14.6	14.6	14.6	14.6	14.6	+1.2	0.0	+0.0	+0.4	+0.4	+0.4	+0.4	+0.6	+0.4
20	5	14	12.5	12.8	13.0	13.6	13.7	13.7	13.7	13.8	+1.5	+1.2	+1.0	+.4	+.3	+.3	+.3	+.4	+.2
2	15	15	15.0	14.3	14.4	14.6	14.6	14.7	14.7	14.6	.0	+.7	+.6	+.4	+.3	+.3	+.3	+.5	+.4
19	1	8	11.0	8.3	8.0	8.7	8.4	8.0	7.9	7.9	−3.0	−.3	.0	−.7	−.4	.0	+.1	.0	+.1
1	20	15	16.8	14.3	14.4	15.0	14.9	15.1	15.1	15.1	−1.8	+.7	+.6	.0	+.1	−.1	−.1	−.2	−.1
1	10	13	13.2	13.0	13.2	13.1	13.3	13.3	13.3	13.1	−.2	.0	−.2	−.1	−.3	−.3	−.3	−.4	−.1
18	8	15	13.5	14.2	14.3	14.5	14.5	14.5	14.5	14.6	+1.5	+.8	+.7	+.5	+.5	+.5	+.5	+.8	+.4
17	4	13	12.0	12.4	12.6	12.6	12.9	13.0	13.0	13.0	+1.0	+.6	+.4	+.4	+.1	.0	.0	.0	.0
2	6	12	11.7	12.2	12.3	12.3	12.6	12.5	12.5	12.3	+.3	−.2	−.3	−.3	−.6	−.5	−.5	−.6	−.3
5	9	14	13.0	14.1	14.2	13.8	13.8	13.9	13.9	13.8	+1.0	−.1	−.2	+.2	+.2	+.1	+.1	+.2	+.2
13	13	15	15.0	15.6	15.4	15.3	15.2	15.3	15.3	15.3	.0	−.6	−.4	−.3	−.2	−.3	−.3	−.6	−.3
8	19	16	16.9	16.0	15.7	15.8	15.6	15.8	15.8	15.9	−.9	.0	+.3	+.2	+.4	+.2	+.2	+.4	+.1
8	12	15	14.4	15.3	15.2	14.7	14.7	14.7	14.7	14.8	+.6	−.3	−.2	+.3	+.3	+.3	+.3	+.5	+.2
15	4	13	11.8	12.5	12.7	12.5	12.8	12.8	12.8	12.9	+1.2	+.5	+.3	+.5	+.2	+.2	+.2	+.2	+.1
3	8	13	12.5	13.3	13.4	13.2	13.4	13.3	13.3	13.2	+.5	−.3	−.4	−.2	−.4	−.3	−.3	−.4	−.2
4	2	9	10.4	9.5	9.5	9.0	8.8	9.2	9.1	8.6	−1.4	−.5	−.5	.0	+.2	−.2	−.1	−.1	+.4
17	17	16	16.7	15.9	15.6	16.0	15.8	16.0	16.0	16.2	−.7	+.1	+.4	.0	+.2	.0	.0	.0	−.2
10	9	14	13.4	14.7	14.7	14.3	14.3	14.3	14.3	14.3	+.6	−.7	−.7	−.3	−.3	−.3	−.3	−.5	−.3
5	13	15	14.5	14.9	14.9	14.6	14.6	14.7	14.7	14.7	+.5	+.1	+.1	+.4	+.4	+.3	+.3	+.5	+.3
10	6	13	12.3	13.7	13.9	13.4	13.5	13.4	13.4	13.4	+.7	−.7	−.9	−.4	−.5	−.4	−.4	−.6	−.4
20	1	8	11.0	8.2	7.9	8.8	8.5	8.0	7.9	8.2	−3.0	−.2	+.1	−.8	−.5	.0	+.1	.0	−.2
1	14	14	14.6	13.7	13.8	14.0	14.1	14.2	14.2	14.1	−.6	+.3	+.2	.0	−.1	−.2	−.2	−.3	−.1
2	17	15	15.8	14.5	14.6	15.0	14.9	15.0	15.0	15.0	−.8	+.5	+.4	.0	+.1	.0	.0	.0	.0
8	10	14	13.6	14.8	14.8	14.4	14.3	14.4	14.4	14.3	+.4	−.8	−.8	−.4	−.3	−.4	−.4	−.6	−.3
12	8	14	13.1	14.5	14.5	14.2	14.2	14.2	14.2	14.2	+.9	−.5	−.5	−.2	−.2	−.2	−.2	−.3	−.2
6	6	13	12.0	13.3	13.5	13.0	13.2	13.1	13.1	12.9	+1.0	−.3	−.5	.0	−.2	−.1	−.1	−.2	+.1
3	12	14	14.0	14.2	14.3	14.2	14.2	14.3	14.3	14.2	.0	−.2	−.3	−.2	−.2	−.3	−.3	−.5	−.2
16	2	11	11.2	10.3	10.4	10.3	10.5	11.0	10.9	10.9	−.2	+.7	+.6	+.7	+.5	.0	+.1	+.1	+.1
10	2	10	10.7	10.3	10.5	9.8	9.9	10.4	10.3	00.1	−.7	−.3	−.5	+.2	+.1	−.4	−.3	−.2	−.1
4	14	15	14.9	14.8	14.8	14.7	14.7	14.7	14.7	14.7	+.1	+.2	+.2	+.3	+.3	+.3	+.3	+.5	+.3
Total		401	401.3	400.6	401.5	400.0	400.5	402.1	401.6	400.7	26.3	12.4	12.3	8.8	8.9	6.9	7.0	------	.63
Alienation coefficient or index		------	.533	.233	.224	.168	.154	.131	.126	.112	------	------	------	------	------	------	------	------	------
Correlation coefficient or index		------	.846	.972	.975	.986	.988	.991	.992	.994	------	------	------	------	------	------	------	------	------
Standard deviation		2.14	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------	------
Standard error		------	------	------	------	------	------	------	------	------	1.14	.50	.48	.36	.33	.28	.27	------	.24

THE CORRELATION ALINEMENT CHART ILLUSTRATED BY AN EXAMPLE

Step 1.—The first step is the determination of the multiple regression equation by the method already described. This has been done and found to be—

$$Z = 0.0635X + 0.3684Y + 9.392$$

The multiple alienation coefficient is 0.54. The corresponding plane surface is illustrated in Figure 9.

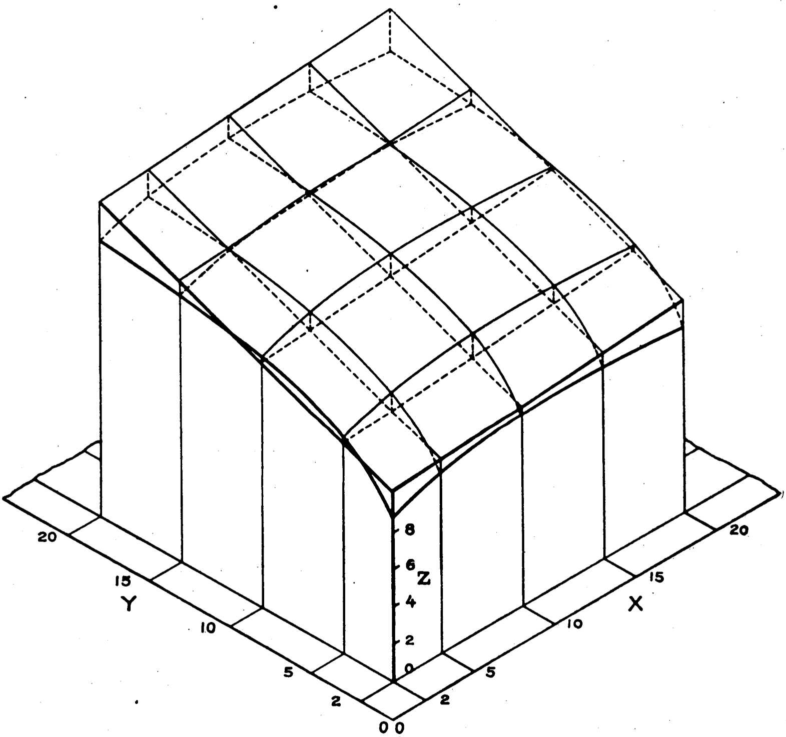

FIGURE 9.—The plane of the linear regression equation (the first approximation) and the curved surface representing the second approximation for the material in Table 6

Step 2.—The second step involves the calculation of estimated values of Z by means of this equation, a decidedly tedious process were there several hundred observations instead of 30. Up to this point in this discussion short-cut methods have not been described, since to do so would merely confuse the reader. In the present instance, however, the use of a time-saving mechanism, the alinement chart, is so intimately associated with the whole procedure that it will be necessary to describe it. An example of an alinement chart which permits rapid computation by the regression equation just given is shown in Figure 10. Its properties are such that any straight line which cuts the three axes will intersect them at values which satisfy this equation. It therefore follows that if values of X and

Y are given, a straight line connecting them will intersect the *Z* axis at the required value. How such charts may be constructed will be described in later pages. For the present it will be assumed that the required chart is at hand. It is used to determine estimated values of *Z*, which are entered in the fourth column of Table 6.

Step 3.—The residuals, or differences between each original (measured) value of *Z* and the corresponding estimated value obtained in step 2 are next computed. These are entered in column 12 of Table 6.

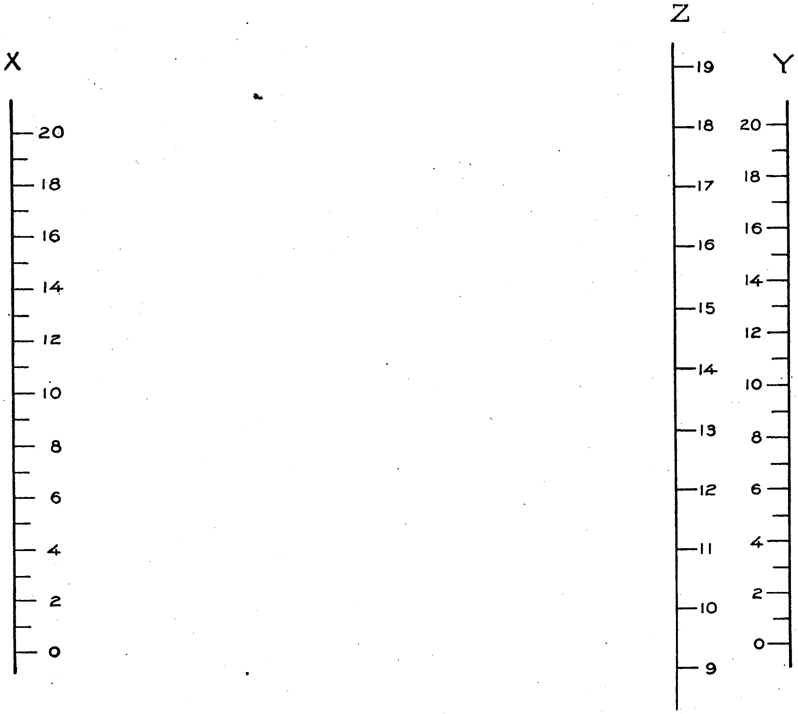

Figure 10.—Alinement chart for the equation (first approximation) $Z=0.0635X+0.3684Y+9.392$. A straightedge, or straight line on a celluloid strip, laid across the chart in such a manner as to intersect the *X* and *Y* axes at any given values of these variables will intersect the *Z* axis at the same value as would be obtained by means of the equation

As a preparatory step for the use of these residuals, the multiple regression equation is next converted into two net regression equations by substituting in it first the mean value for *Y* and then the mean value for *X*. These substitutions give—

$$Z=0.0635X+12.79$$

$$Z=0.3684Y+9.97$$

Lines corresponding to these two equations are then drawn, the straight lines of Figure 11. The residuals are next plotted about the net regression line for *X*, their horizontal position being determined by the corresponding *X* values and their vertical positions being measured, not from the horizontal axis, but from the regres-

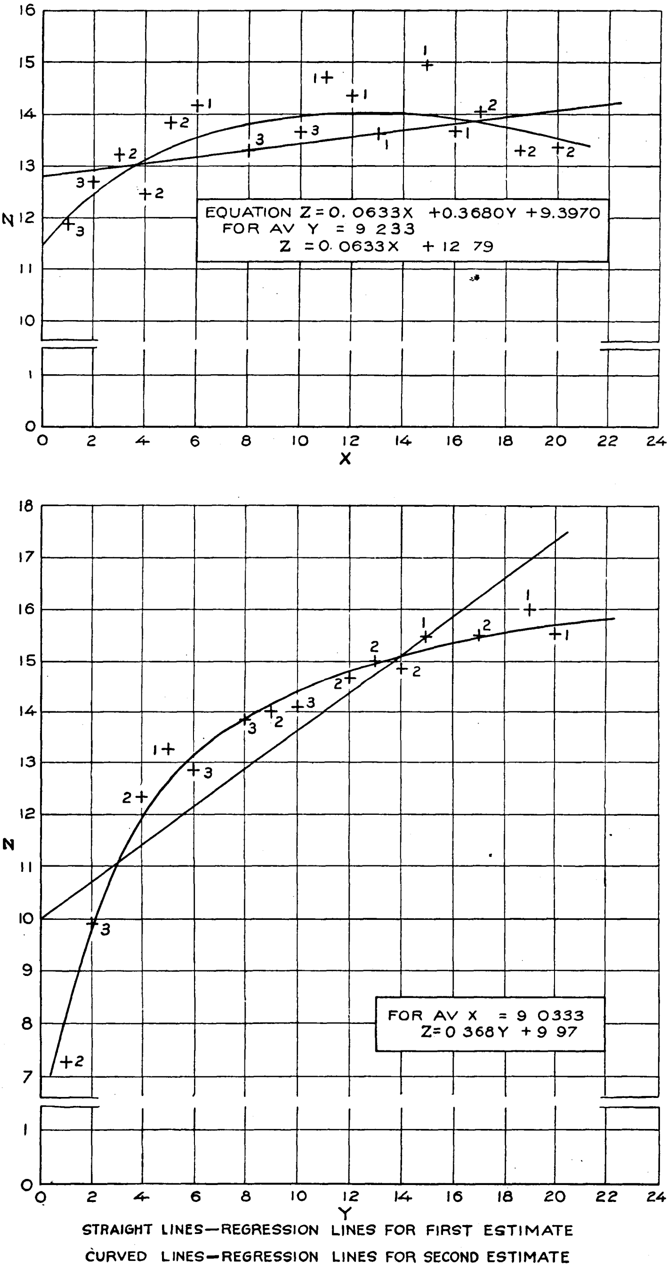

FIGURE 11.—The regression straight lines and first approximate regression curves for the data in Table 6. The straight lines fit the data poorly

sion line, above or below, according to their signs. (Class averages rather than individual values are shown in fig. 11.) The same process is then repeated for the net regression line for Y. Free-hand curves are then fitted to each series of points without difficulty. In both cases there are decided indications of curvature.

This process is analogous to that outlined on page 14 and illustrated in Figure 4, except that the residuals are plotted about the regression line instead of about a zero horizontal axis. Had the regression equation been an adequate expression of the data, there would have been no correlation between the residuals and either of the independent variables. The points plotted by means of the residuals (as in fig. 11) would then define the regression lines. That this is not the case in the present instance is evidence of the existence of curvilinearity.

Step 4.—These curves are next used to calculate the second estimated values of Z (column 5, Table 6). This may be done by readjusting the graduations of the X and Y axes of the alinement chart to make them agree with these curves instead of the straight lines from the regression equations. For example, in the lower part of Figure 11 it will be seen that the value from the curve corresponding to $Y=20$ is the same as that from the regression straight line for $Y=15.5$. The revised 20 graduation is, therefore, placed where the 15.5 graduation was originally located. In a similar way 19 is placed where 15.3 was originally, etc., etc. When both X and Y scales of the alinement chart are thus completely revised (fig. 12), the second estimated values for Z may be read from it directly.

Although the second estimate residuals are not used, they are entered in column 13 for the purpose of comparison.

Figure 9 shows how the plane of the regression equation has in effect been modified by this treatment. The curved surface is of the type shown in Figure 8, B.

Step 5.—It is at this point that the possibility of a functional relationship between Z and $f_1 (X) + f_2 (Y)$ may be investigated. In correcting the alinement chart, variably spaced graduations have been substituted for those of uniform interval on the X and Y scales. The possibility of an improvement through a similar transformation of the Z scale is obvious. Such a transformation may be readily accomplished as follows:

For convenience, the measured and second estimate values of Z are first sorted into classes, the sorting basis being the second estimates, as illustrated in Table 7.

TABLE 7.—*Example of sorting of second estimate and measured values on basis of second estimates, in step 5*

Second-estimate class	Number of items	Average value of Z		Second-estimate class	Number of items	Average value of Z	
		Second estimate	Measured			Second estimate	Measured
8.0 to 8.9	2	8.25	8.0	13.0 to 13.9	5	13.4	13.2
9.0 to 9.9	1	9.5	9.0	14.0 to 14.9	11	14.5	14.5
10.0 to 10.9	2	10.3	10.5	15.0 to 15.9	4	15.4	15.2
11.0 to 11.9	0	----------	----------	16.0 to 16.9	1	16.0	16.0
12.0 to 12.9	4	12.5	13.0				

The desired correction is to place the 8.0 graduation where the 8.25 graduation (which is not actually marked) now is, and so on. The various corrections indicated must be harmonized, however, and this is accomplished by plotting the average measured values of this table over the average second estimates and fitting a curve to them. (Fig. 13.) By means of this curve the Z axis may be regraduated. For example, reading the curve backwards, the revised 8.0 graduation is placed where the original 8.3 graduation was located.

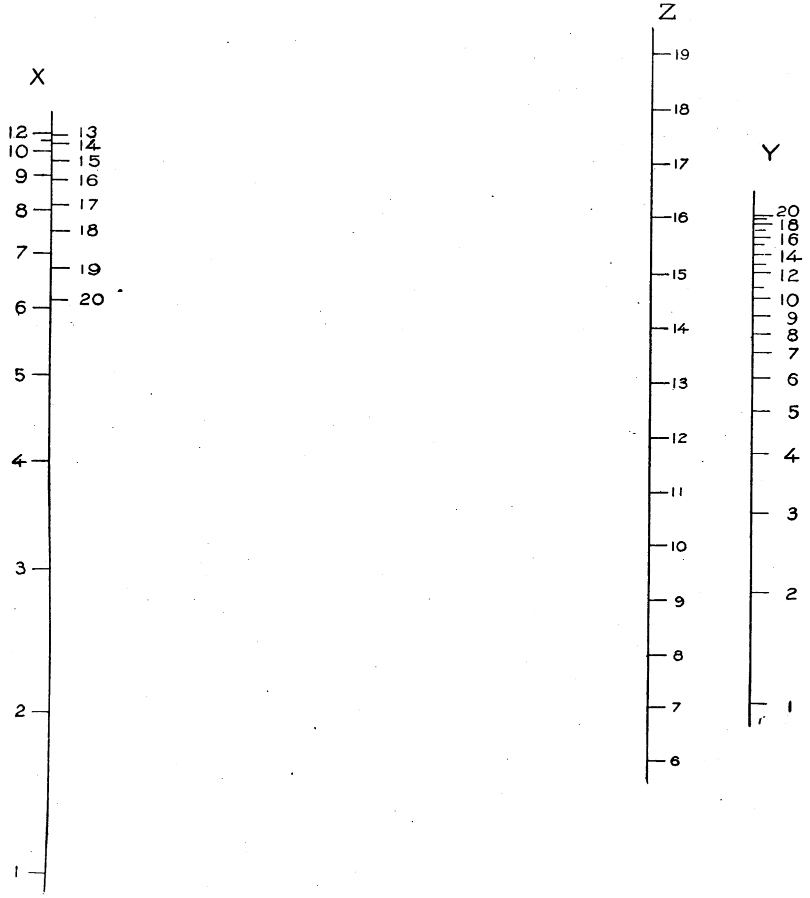

FIGURE 12.—Alinement chart for the second estimation of values of Z from values of X and Y. This is derived from Figure 10 by means of the curves of Figure 11. The folding back of the graduations of X is the result of the rising and falling curve for that variable in Figure 11

In the present instance, the curvature, though well-defined, is slight, and hence may be accidental. It would not be at all surprising to see it eliminated in later stages. As will be seen, however, it becomes more and more accentuated.

With the Z axis of the alinement chart regraduated to correspond to this curve, the third estimates of Z are obtained.[10] (Column 6, Table 6.)

[10] If preferred, these third estimates may be read directly from Figure 13 (the curved value of measured Z corresponding to the second estimate becomes the third estimate), but the alinement chart must be revised in any event for use in subsequent steps.

This step will be recognized as similar to that described on page 15 and illustrated in Figure 5. Had no correction been required, the points of Figure 13 would have defined the 45-degree straight line passing through the origin. Had they defined a straight line not passing through the origin, this would have indicated slight errors in fitting the curves of Figure 11.

Step 6.—The next process is a repetition of the third and fourth steps and is illustrated by Figure 14 and columns 14, 7, and 15 of Table 6. The residuals of this estimate are plotted about the curves of Figure 11 instead of about the regression straight lines. It will be seen that in both portions of Figure 14, the curvature has been reduced. In regraduating the axes of the alinement chart to correspond to the revised curves, it is somewhat easier to refer the new graduations to the uniformly spaced graduations of Figure 10, and for this reason the original regression lines (dotted) are entered in Figure 14.

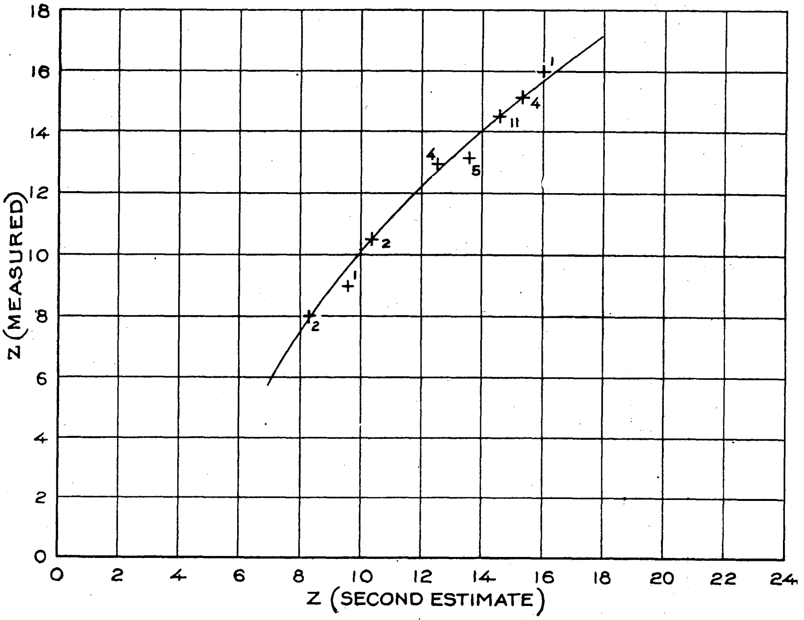

FIGURE 13.—The average measured values of Table 7, plotted over the average second-estimate values. The ourve fitted is used for revising the *Z* axis of Figure 12. It can also be used for reading a third estimate from the second-estimate values

Step 7.—The next step is a repetition of the fifth and is illustrated by Figure 15 and column 8 of Table 6. The curvature suggested by Figure 13 is here more strongly defined.

Subsequent steps.—In a similar manner those two types of successive approximations may be alternately applied until it is is seen that no further improvement is being made. The final graph of the type of Figures 13 and 15 should, of course, be approximately a 45-degree straight line, and this is illustrated in Figure 16.

Final step.—In some cases, particularly where the independent variables are closely correlated with each other, the successive curves

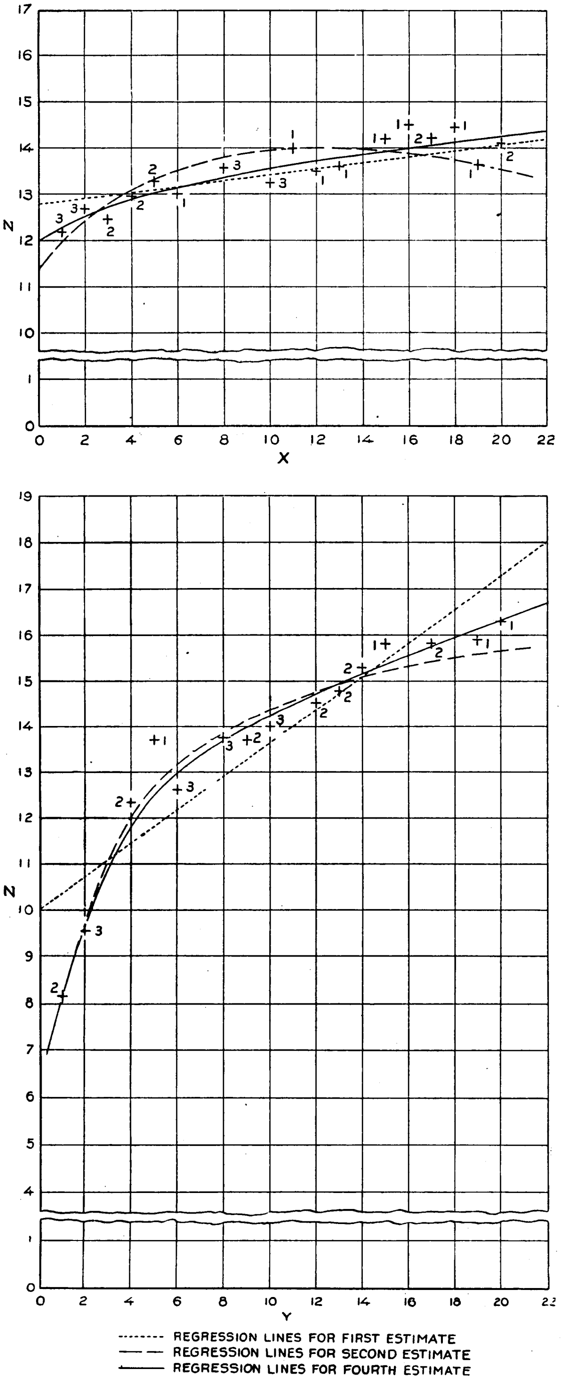

FIGURE 14.—The residuals of the third estimate, column 14, Table 6 (plotted about the second-regression lines), to which the regression lines for the fourth estimate are fitted. The curvature is less in each case. This is not uncommon where the two lines curve in the same direction

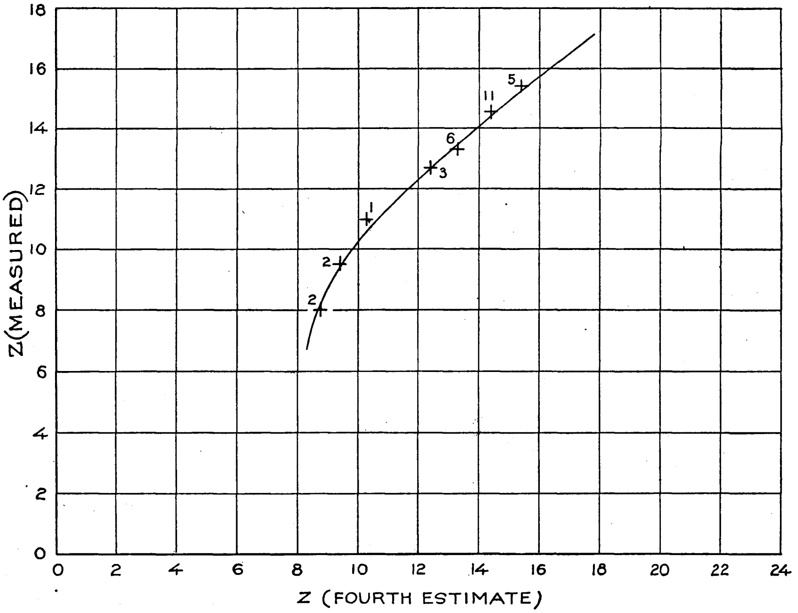

FIGURE 15.—Average measured values plotted over the averages of the fourth-estimate values, preparatory to again revising the *Z* axis and making the fifth estimate. The slight curvature indicated by Figure 13 is not only confirmed but increased

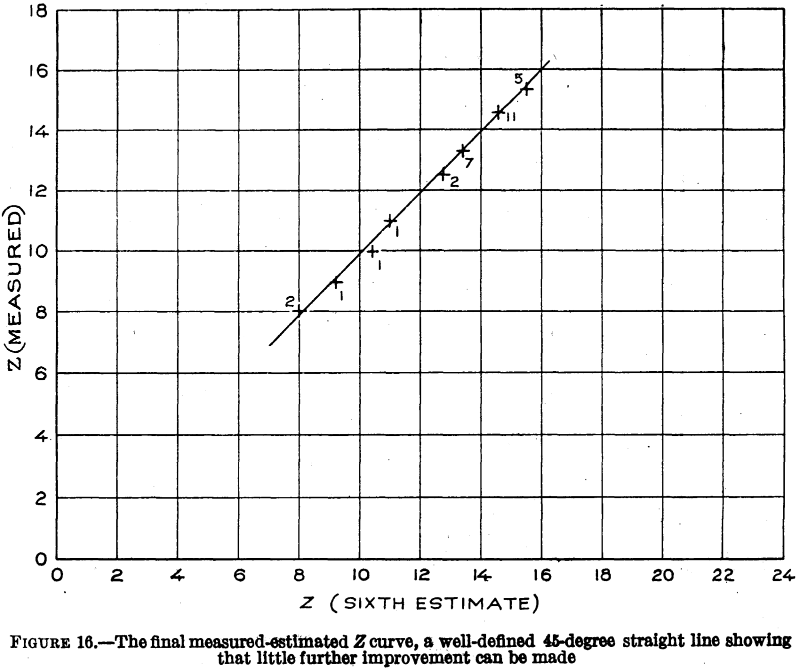

FIGURE 16.—The final measured-estimated *Z* curve, a well-defined 45-degree straight line showing that little further improvement can be made

may be found to swing back and forth instead of settling down into well-defined and stable positions. In such cases the final correction may be made by calculating a multiple regression equation between the last residuals and the independent variables, such as—

$$e=AX+BY+C$$

where e is the residual of the last estimate.

Had the previous steps been so completely successful that further corrections were impossible, this would be evidenced by zero values for A, B, and C. If they were not equal to zero it would be clear that the part of the residuals which is associated with X and Y may be eliminated by appropriate corrections, and the estimates further improved thereby.

Although there is no evidence in the curves that such a correction is desirable in the present instance, it is interesting to perform it, if only to illustrate the method. If a regression equation be computed from the values of column 18A and columns 1 and 2, the following results—

$$e=0.0227X+0.0185Y-0.396.$$

The alienation coefficient is 0.884. Obviously, the small corrections implied by this equation are insignificant. An accurate appraisal of their importance can best be made by a calculation of the following type. The alienation index obtained by the seventh estimate is 0.126. The alienation index resulting from applying this correction will be the product $0.126\times0.884=0.111$. It is hard to conceive any practical case in which an improvement of this magnitude, 0.126 to 0.111, would justify the labor of making it, but in the present instance it will be carried out in order to illustrate the method.

The correction equation can readily be expressed by an alinement chart, but unfortunately it can not be used to correct the existing chart except in cases where the dependent variable has not taken on a functional form. This would be awkward, since an additional chart would be needed for the final correction. The difficulty of combining two such charts comes from the fact that the correction equation is of the form

$$e_Z=AX+BY+C$$

and this can not be added to

$$f_0(Z)=f_1(X)+f_2(Y)$$

so as to yield an equation of similar form. It would, however, be possible to add a correction equation of the type

$$e_{f_0(Z)}=AX+BY+C$$

or this would result in

$$[\text{corrected}\,f_0(Z)]=[f_0(Z)+e_{f_0(Z)}]=[f_1(X)+AX]+[f_2(Y)+BY]+C.$$

Interpreted graphically this difficulty is associated with the divergence and convergence of the graduations of the Z axis of the alinement chart, in terms of which the residuals are expressed.

To express the residuals in terms of $f_0(Z)$ all that is necessary is to add to this axis a convenient auxiliary scale with uniform spacing.[11] The multiple regression equation between these residuals and the independent variable is computed and substituted for that just described.

Figure 17 shows the alinement chart by means of which the seventh approximations are computed, with this auxiliary scale added to the Z axis. The residuals computed thereby are given in Table 6 in column 18, B. The multiple regression equation between these and the independent variables is

$$e_{f_0(z)}=0.0272\ X+0.0255\ Y-0.508$$

The coefficient of alienation is 0.925, which differs but little from that already obtained (0.884) for the equation for e_Z. Modification of the alinement chart for the seventh estimate to provide for this correction involves raising the graduations of the X axis sufficiently to increase the final readings by 0.0272 X, raising those of the Y axis sufficiently to increase the final readings by 0.0255 Y, and finally raising those of the Z axis by a uniform 0.508 so that the final readings will be reduced by this constant. To produce this result[12] the X graduations must be raised—

$$0.0272\ X\left(\frac{\text{distance between } X \text{ and } Y \text{ axes}}{\text{distance between } Z \text{ and } Y \text{ axes}}\right)=0.0272\ X\left(\frac{6.8}{1.0}\right)=0.185\ X$$

the result thus calculated being in terms of the auxiliary units added to the Z axis. In a similar way the Y graduations must be raised—

$$0.0255\ Y\left(\frac{\text{distance between } X \text{ and } Y \text{ axes}}{\text{distance between } Z \text{ and } X \text{ axes}}\right)=0.0255\ Y\left(\frac{6.8}{5.8}\right)=0.0299\ Y.$$

Moreover, instead of raising the Z axis, it is somewhat easier to drop both the other axes the same amount, thus making the correction equations:

$$e_{f_0(z)}=0.185\ X-0.508$$

and

$$e_{f_0(z)}=0.0299\ Y-0.508$$

The shifted graduations are shown in Figure 17, and the final estimates and their residuals are given in Table 6, columns 11 and 19.

This type of final correction is not as laborious as might appear at first sight. The intercorrelations between the independent variables have previously been calculated in connection with the first step, which materially shortens the labor involved. In most cases the alienation coefficient of the resulting regression equation will indicate that the application of the correction is not justified.

The progressive improvements which result from the successive approximations are indicated by the values of the standard deviation and standard errors and of the alienation and correlation indices

[11] The absolute distance between the measured and last estimated values on the Z axis is measured by this auxiliary scale, or, more simply, the difference is computed between the auxiliary scale values opposite each measured value and its corresponding last estimate. These are the residuals expressed in terms oj the arbitrary scale.

[12] See page 50 for a more complete discussion of this correction.

computed therefrom, entered at the bottom of Table 6. It will be seen that the final standard error and the final alienation index are exceedingly low, particularly when it is remembered that in setting up the measured values the calculations were carried out but to the nearest unit.

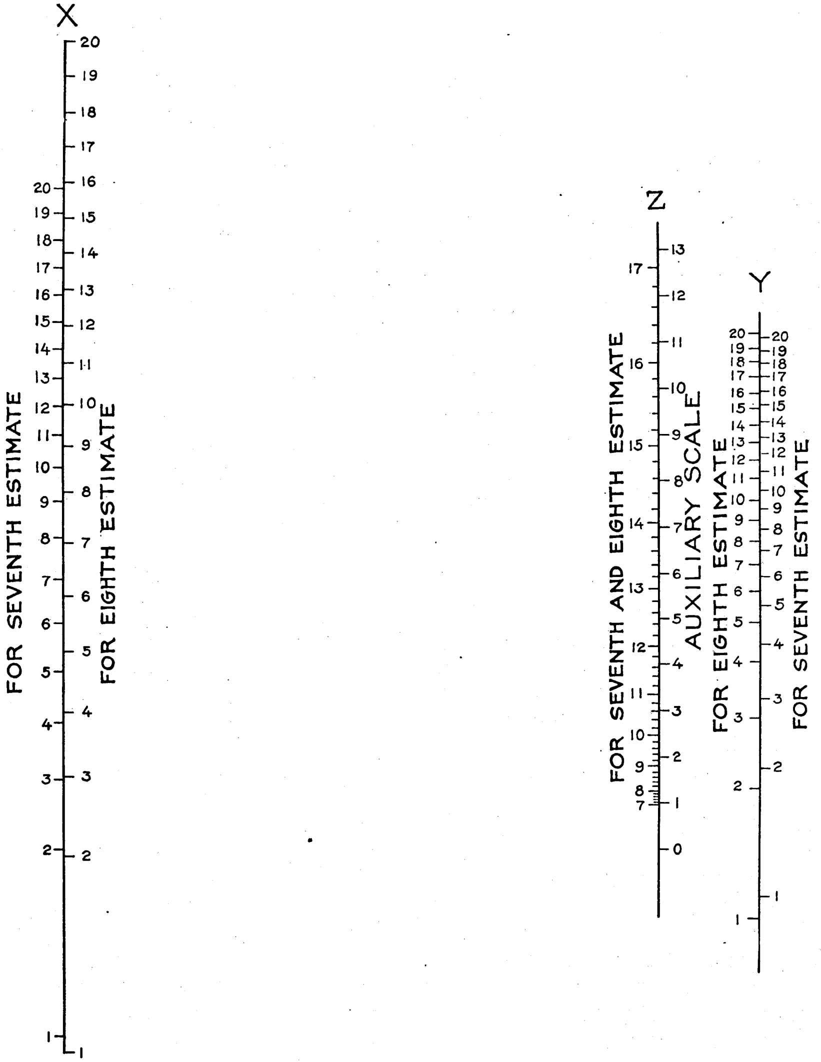

FIGURE 17.—Final alinement chart used for computing the seventh and eighth estimates, showing an auxiliary scale used for measuring the seventh residuals in terms of f_0 (Z) for use in the final correction

The sums of columns 4 to 11, inclusive, constitute a valuable check on the accuracy of the work as it progresses. Each sum should equal the sum of column 3. Minor differences, such as appear in the present example, may be attributed to slight inaccuracies in curve fitting or the like, but any discrepancies too great to be explained in such ways indicate some serious error either in computations or in graphic work. It should also be noted that the difference between each sum and that of column 3 should equal the algebraic sum of the corresponding column of residuals. It is, therefore, worth while to add, algebraically, columns 12, 13, 14, 15, and 16, etc., to permit this additional check.

It remains to be seen whether the curvilinear functional relationships which have been empirically determined are those of the assumed equation used in setting up the problem. Figure 18 shows in isometric projection both the surface of the original equation and that of the final alinement chart. The correspondence is as nearly perfect as could be expected with the limited data used.

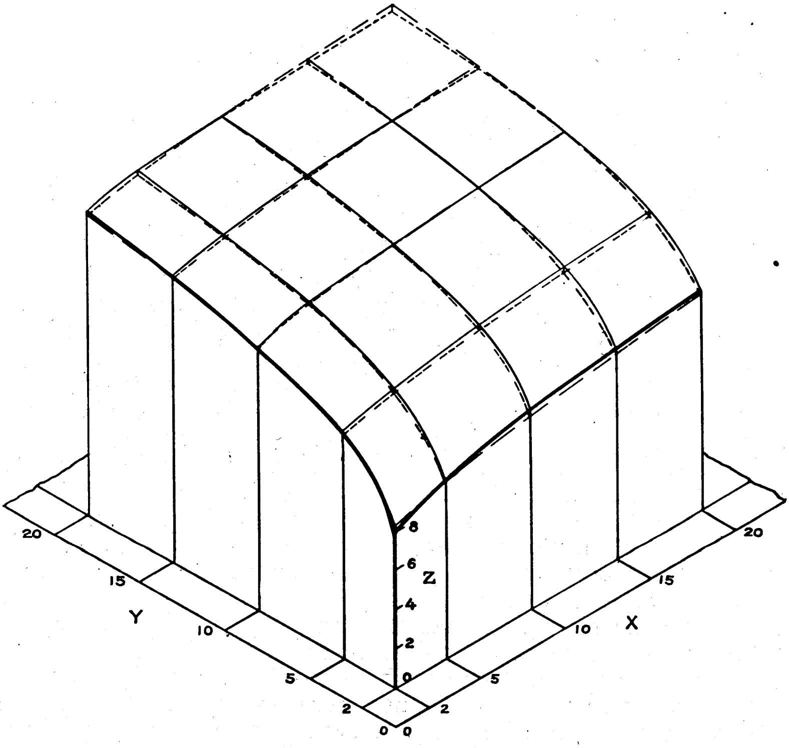

FIGURE 18.—The close agreement between the equation from which data were computed and the estimates from the final chart is shown by the surfaces representing them. Solid lines represent the equation

Certain minor modifications of the general plan herein outlined suggest themselves. For example, it might in some instances be preferable to reverse the order of applying the two different graphic processes. This would be particularly useful in cases where the dependent variable has a curvilinear functional form as, for example, where the basic equation is such as—

$$Z^2 = 3X - \frac{Y}{7}$$

Here a correct result would obviously be approached far more rapidly were the functional relations of Z tested first. Unfortunately it will probably be but rarely that sufficient advance knowledge is at hand to permit an assured determination of which order will be best.

It should be noted that equations involving a multiplication of functions such as—

$$Z=f_1\ (X)\,f_2\ (Y)$$

can be handled by the use of logarithms through conversion into—

$$\log Z=\log f_1\ (X)+\log f_2\ (Y)$$

In a similar manner—

$$Z=f_1\ (X)^{f_2(Y)}$$

can be converted first into—

$$\log Z=f_2\ (Y)\log f_1\ (X)$$

and then into—

$$\log\log Z=\log f_2\ (Y)+\log\log f_1\ (X).$$

With other methods, lack of advance knowledge as to the form of the equation involved usually prevents taking full advantage of this fact. It is always possible, of course, to try the effect on the alienation index of the logarithmic treatment, but the amount of labor involved is excessive. It will be noted, however, that in the converted forms the preceding equations are of the type assumed by the present technic. Since this technic permits a functional form of Z as well as of X and Y, the preliminary conversion into logarithms or log logarithms is not necessary. It therefore follows that such equations can be handled without special treatment, and without preliminary knowledge that they are being encountered.

It should be emphasized, however, that wherever a preliminary analysis of a problem indicates that the logarithmic conversion is suitable, there will be a material economy of time and labor if it be performed. If not, the technic which has been described will ultimately arrive at the same result, but more slowly. The fact that all the functions are logarithmic has a tendency to make convergence towards the true curves abnormally gradual. Cases of this type have been encountered where 17 approximations have been necessary.

Problems involving more than three variables, although somewhat more tedious, present no new complexities. Alinement charts may readily be constructed for four or more variables, and the successive steps are the same as those already described.

In later pages several illustrative cases will be presented in which forestry problems will be worked out. In connection with each of these one or more variations in technic will be described which have not previously been mentioned. Before taking up these actual examples, however, it will be necessary to digress and explain more completely how the alinement charts needed may be prepared.

CONSTRUCTION OF THE ALINEMENT CHARTS NEEDED

This discussion will be restricted to the design of that type of alinement chart needed to express multiple regression equations. It is unnecessary to explain the underlying principle[13] or to discuss other types, but it is desirable to present two methods applicable to multiple correlation problems.

The chart is best made on ordinary cross-section paper, for this gives an easy basis for the drawing of as many parallel axes as may be needed, in any desired position and already graduated at uniform intervals.

The simplest type of multiple regression equation is

$$Z=AX+BY+C$$

In this case the two edges of the graph may be adopted as the initial or X and Y axes. It is usually convenient to disregard temporarily the values A and B, and arbitrarily to use any convenient scale for graduating these two axes. The only precaution necessary is that a sufficient range of values be entered on each, and that they shall not be too crowded together. It is wise also to leave some space at both the upper and lower end to take care of possible shiftings and expansions which may become necessary in subsequent stages of the work.

From the form of the equation it is known that the Z or Sum axis will be a straight line located somewhere between the other axes and parallel to them, and that its graduations will be uniform. A convenient method of determining, by intersection, its position and the size of the graduating interval, is illustrated by the following example. Let us assume that the equation in question is

$$Z=0.5X+0.7Y+4$$

and that we have graduated the outer axes as illustrated in Figure 19. The next step is to find two pairs of values for X and Y yielding identical values of Z; for example, if Z equals 10, this equation becomes—

$$10=0.5X+0.7Y+4$$

or,

$$0.5X+0.7Y=6$$

Now if

$$X=0, \text{ then—}$$

$$Y=\frac{6}{0.7}=8.57$$

and if

$$X=10\text{—}$$

$$Y=\frac{6-5}{0.7}=1.43.$$

Two straight lines, one connecting the point $X=0$ with the point $Y=8.57$ and the other connecting the point $X=10$ with the point $Y=1.43$, must both cut the Z axis at the 10 graduation. They are drawn as indicated by the broken lines AB and CD in Figure 19. Their point of intersection must fall on the Z axis, which may, therefore, be drawn through this point parallel to the other axes.

[13] For a more complete discussion, see Lipka (*24, ch. 3 to 5*) or Peddle (*36*).

In a similar way, it may be assumed that $Z=5$. The equation now becomes—

$$5=0.5X+0.7Y+4,$$

or,

$$0.5X+0.7Y=1$$

Now, if

$$X=0, \text{ then—}$$

$$Y=\frac{1}{0.7}=1.43$$

and if

$$X=5\text{—}$$

$$Y=\frac{1.5}{0.7}=-2.14$$

The corresponding lines are drawn; these are AD and EF. The fact that their intersection lies on the axis attests the correctness of its position. Furthermore, the distance between the two intersection

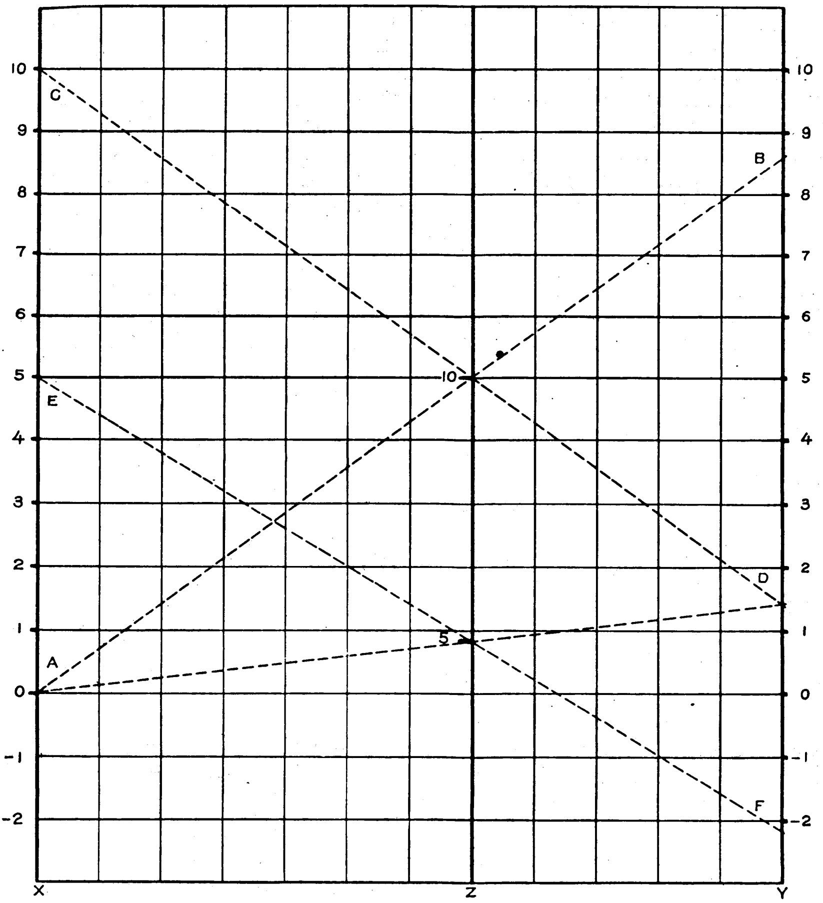

FIGURE 19.—This illustrates the construction, by intersections, of an alinement chart based on the formula $Z=0.5X+0.7Y+4$

points must obviously represent $10-5$, or 5 units on the new scale. One-fifth of this distance may then be determined and used to locate the remaining graduations.

Where the coefficient of X is much larger than the coefficient of Y, the Z axis will fall very close to the X axis. If this is the case it is somewhat more desirable to use a smaller scale in graduating the Y axis than is used for the X axis, as this will shift the Z axis toward a more central and hence a more convenient location.

A slightly different situation presents itself where one of the regression coefficients is negative; as, for example, in an equation such as—

$$Z=0.5X-0.7Y+4.$$

This can be handled by the simple expedient of graduating Y in the opposite direction from X, i. e., from top to bottom, and if this is done, the intermediate or Sum axis falls in the same position (and has the same graduating unit) as it would were both coefficients positive. There is some danger of erroneous readings where this is done, however, and a preferable plan is to so transpose the terms of the equation as to eliminate the minus sign of the variable Y. For example, the equation can be written—

$$0.5X=Z+0.7Y-4.$$

The construction now proceeds as before, except that the X axis instead of the Z will now be in the central position.

Where both coefficients are negative as, for example, in the equation—

$$Z=-0.5X-0.7Y+4$$

such a transformation is impracticable, and both X and Y axes must be graduated in a direction opposite to that used for the Z axis.

The alinement chart for four or more variables appears more complicated, but no new principle is involved. Take, for an example—

$$W=0.5X+0.7Y+0.6Z+2.$$

The method is to split this equation into two parts by assuming—

$$S=0.5X+0.7Y \text{-------------------} (A)$$

whence—

$$W=S+0.6Z+2 \text{-------------------} (B)$$

An alinement chart for (A) is prepared as before. In Figure 20 this is represented by the axes X, Y, and S. Using this S axis as a starting point, a second chart is now prepared and superposed on the first, the Z axis being arbitrarily located, and the position and graduating interval of the W axis being determined by intersections based on equation (B).

In using this chart a straightedge is first laid across the values for X and Y to determine the value of S. It is then shifted to connect this value of S with that given for Z, and the required answer is then read from the W axis. The process is illustrated by the dotted lines of Figure 20 by means of which the W value 8.3 is obtained when $X=2$, $Y=5$, and $Z=3$.

The sequence in which the axes are used can not be changed. The proper sequence is, therefore, usually expressed by a key which gives the specific order in which to perform the various shifts of the straightedge. For this chart the key would be:

> From (a point on) *X* (go) to (a point on) *Y*, hold (the intersection) on *S*; (go) to (a point on) *Z*, read *W*.

The parenthetical expressions are usually omitted, and the above key would be expressed:

> From *X* to *Y*, hold *S*; to *Z*, read *W*.

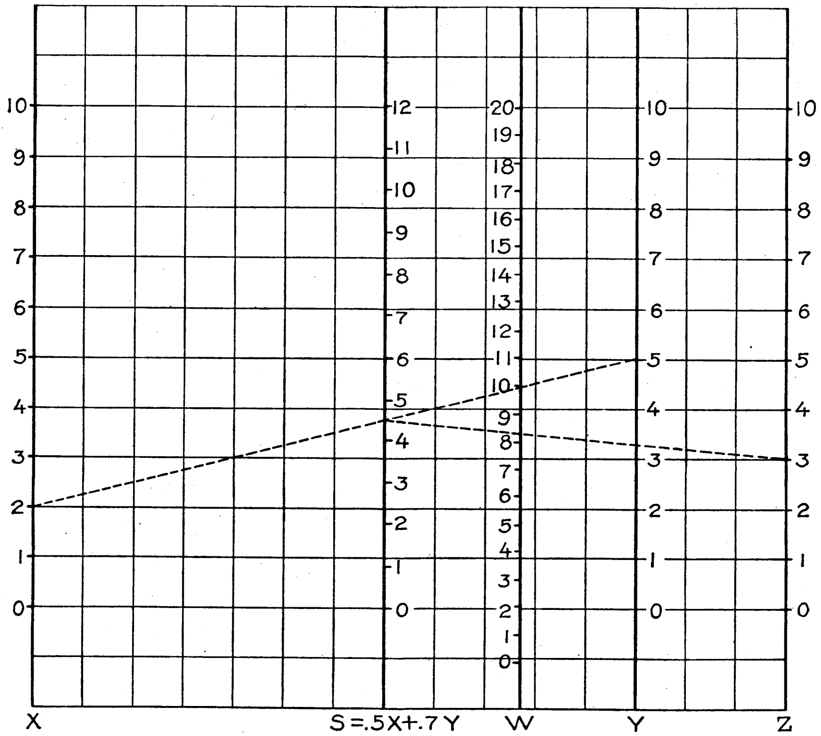

FIGURE 20.—This alinement chart expresses the 4-variable equation $W=0.5X+0.7Y+0.6Z+2$. The graduations on the *S* axis, while useful in constructing the chart, may later be erased, as no readings are made on this axis

In reading these charts it is convenient to mount them on a drawing board and then use needles (with sealing wax heads to facilitate handling) to hold the points read. If this be done, the graduations may be erased from the *S* axis, for the *S* values need not be actually read, their positions on this axis being held with a needle while the straightedge is shifted.

In an analogous manner charts for five or more variables may be prepared. Each additional variable requires two more axes and one more shift of the straightedge. While the labor in using these multivariable charts is great as compared with the 3-variable ones, it is small when compared with that involved in computing by means of the equation.

In charts for equations with several variables, however, this graphic method of locating the positions and graduations of the axes by intersections is not entirely satisfactory. Graphic inaccuracies may accumulate in the successive stages until the final axis is materially in error. It is preferable in such cases to determine the positions and scale units by computations rather than by graphics.

The term "modulus" has been adopted for the constant which is used in graduating any axis. It may be defined by the equation

$$U=Lf(\ )\text{-----------------------}(X$$

where L is the modulus of any variable, U is the distance of any graduation from the zero point on the corresponding axis, and $f(\ )$ is the function of the variable.

To illustrate, in the case just solved graphically (p. 37) where the equation might be written—

$$(Z-4)=0.5X+0.7Y$$

the modulus of the X axis is given by the equation—

$$U_X=L_X(0.5X)$$

In Figure 19 it will be seen by inspection that where $X=1$, $U_X=$ one division of the graph paper above $X=0$. Therefore—

$$1=L_X(0.5)$$

and

$$L_X=2$$

The equation of the modulus of the Y axis is—

$$U_Y=L_Y(0.7Y)$$

Since, by inspection of Figure 19, where $Y=1$, $U_Y=$ one division of the graph paper above $Y=0$. Therefore—

$$1=L_Y(0.7)$$

and

$$L_Y=\frac{1}{0.7}=1.429$$

From this the two graduating equations may be written,

$$U_X=2(0.5X)$$

or

$$U_X=X$$

and

$$U_Y=1.429(0.7Y)$$

or

$$U_Y=Y$$

This merely illustrates the use of the moduli, and serves no useful purpose in so simple a case, in which the scale was arbitrarily made to coincide with the ruling of the coordinate paper.

However, if L_Z be the modulus of the Z axis, it may be determined by the formula—

$$L_Z=\frac{L_X L_Y}{L_X+L_Y} \text{ ------------------- (XI)}$$

In the present instance—

$$L_Z=\frac{2\times 1.429}{2+1.429}=\frac{2.858}{3.429}=0.833$$

and the graduating equation for the Z axis is—

$$\begin{aligned} U_Z&=L_Z f(Z) \\ &=0.833(Z-4) \end{aligned}$$

It will be found that the Z axis of Figure 19 can readily be graduated by means of this equation if the 0 point of U_Z thereon be so chosen as to fall on the straight line connecting the 0 points of the two other scales.

The position of the sum axis is, moreover, defined by the equation—

$$\frac{XZ}{ZY}=\frac{L_X}{L_Y} \text{ -------------------- (XII)}$$

where XZ and ZY are the distances from X to Z and Z to Y, respectively. The distances are from the left-hand initial axis to the sum axis and from the latter to the other initial axis; distances from left to right are considered positive. In the present instance—

$$\frac{XZ}{ZY}=\frac{2}{1.429}$$

Furthermore, the X and Y axes have been arbitrarily placed 12 units apart so that—

$$XY=XZ+ZY=12$$

whence

$$XZ=12-ZY$$

We may, therefore, write—

$$\frac{12-ZY}{ZY}=\frac{2}{1.429}$$

whence

$$ZY=5.00$$

and

$$XZ=12-5.00=7.00$$

These two distances will be seen to agree with those obtained graphically in Figure 19.

It will be seen that when any convenient assumptions are made as to positions and moduli of any two of the axes, the position and modulus of the third may be rigorously calculated by means of Equations X and XI. In equations involving several variables, where this algebraic method is always preferable to the graphic, any desired assumptions may be made concerning the first two axes, such as X and Y in Figure 20. The position and modulus of the cumulating or Sum axis, S, is then calculated by the formulæ. A new assumption may be made then as to both the position and scale of either the W or Z axis, and similar information calculated by means of the same formulæ for the other. In this method the actual graduations need never be entered on the S axis.

The positions of the scales on the axes can be shifted up or down as desired, whether the graduations are regular or not. The only point to keep in mind is that any set of values must give the same answer after shifting as they gave before. This property of straight axis charts is of value sometimes where a scale is greatly expanded at one end and would go off the paper if the entire scale was not shifted. Any two scales may be placed arbitrarily in convenient positions, and the requisite shift of the remaining scales, necessary to compensate for it, can then be determined by intersections.

PRELIMINARY ANALYSIS OF THE PROBLEM

Now that the use of alinement charts in correlation problems has been explained, a series of examples will be presented to illustrate how the method works when applied to actual forestry problems. If such problems are to be solved with the least effort and the greatest accuracy a thorough preliminary analysis is essential before the actual correlation calculations are undertaken. The chief objective must be decided on, and all factors entering into the problem should be considered so that those of minor importance may be determined and rejected when such rejection becomes desirable.

The primary purpose of any problem to which this correlation method is applied will be (1) to define the quantitative and qualitative relations between a given factor and the factors affecting it; (2) to produce a mechanism for estimating values of one variable from measurements of variables associated with it; or (3) to accomplish both of these results.

In the first class of problems, where the chief objective is to define and measure the relation between a given effect and its causative factors, it is necessary to consider all causative factors, separately or in groups. For example, the height growth of a stand of trees is the direct result of duration of growth, soil and climatic conditions, inherent characteristics of the trees, condition and vitality of the individuals and of the stand as a whole, and of various abnormal influences such as fire and attacks of insects and fungi. Here the climatic conditions may be treated separately, precipitation, evaporation, insolation, temperature, etc., each considered by itself, or the soil and climatic conditions may be treated en masse as site quality.

In some sciences the effect of any one of the variables can be investigated by keeping all others constant. In such cases, of course, the results apply only to the particular set of conditions under which the investigation was made. Variation in the dependent variable may be due to the joint effect of two or more causative factors, and a repetition of the investigation must be made for other combinations of the values of those causative factors.

In the great majority of forestry problems it is impossible to keep constant or to control any factor or set of factors, and in most cases there are some which can not be isolated and measured directly.

Their combined effect must be measured and treated as a single factor. Certain abnormal factors, such as damage from fire, insects, and fungi, can be eliminated by a proper choice of samples.

In the second class of problems, where the chief object is a mechanism for prediction, controlled experiments are usually impossible, and a limited number of variables must be used. In the case just given (height growth) a sufficiently accurate estimate could probably be made from an evaluation of two of its causes, age and site quality. The inclusion of more variables might result in too small an increase in accuracy to warrant the greater complexity of the predicting mechanism or the greater expenditure of labor necessary for its construction.

The predicting mechanism just cited is based on measurements of the causative factors. In other instances an estimate of the result of certain causes may be made from measurements either of the immediate causes or of other factors less directly related to that result, as in the use of diameter and height to estimate defect in trees. In this case the defect, while not caused by either height or diameter growth, may be due in part to factors influencing both. In still other cases the variable estimated may be in part a mathematically exact function of the variables used in its prediction, as in the estimation of tree volume from measurements of diameter and height.

In either class of problem the dependent variable must be decided upon in advance so that the alinement charts can be properly constructed. The variables in such charts can not legitimately be reversed any more than in other graphic processes. The errors resulting from such a reversal are particularly serious when the alienation index is high.

COLLECTION OF DATA

Once the objectives of the project have been decided on and the factors entering into it have been enumerated, the necessary data should be collected in proper form for the method of treatment decided upon.[14]

Since the correlation method is founded on mathematical principles it is necessary to evaluate all factors involved in quantitative units, rather than in qualitative terms. Of the three major variables involved in most forestry problems (1) duration, or time, (2) heredity, and (3) the composite site quality, the first is easily evaluated in the customary units of time, while heredity can hardly be measured and, therefore, must usually be eliminated from direct consideration. The composite site quality must be evaluated in commensurable units [15] rather than by the customary relative classification as good, medium, or poor. Of the many subfactors entering into site some can be evaluated directly, while others must be measured through their effects as evidenced by size, form, strength, etc. The climatic factors entering into site are measurable, but complete and reliable records are costly to obtain and seldom available, so that measurements of climate must usually be obtained through measurements of its effects. Soil quality and competition can seldom be measured numerically,

[14] In this connection it will be well worth while to read Day (*8, ch. 2 and 3*). Although the viewpoint is that of business statistics, the text in general is valuable.

[15] Quality of site can be expressed in terms of site index. Site index is the height attained, at a given age, by the average dominant tree growing on the area.

and they, too, must be measured through their effects. Site, therefore, must usually be evaluated as a composite factor, in terms of a site index.

In the indirect measurement of causes, as in their direct measurement, quantitative units must be employed instead of qualitative terms. In the direct measurement of insolation the units, gram-calories, should be used; so, also, in measuring it indirectly through aspect the measurement of aspect must be in degrees of azimuth rather than in the descriptive terms of the points of the compass, such as NE, SE, etc.

In certain problems consideration must sometimes be given to incommensurable factors such as locality. This can best be done by first treating all localities together and, as a subsequent step, analyzing the deviations from the composite of the material obtained in each locality.

In other problems certain established facts can be legitimately employed in their solution without incorporation into the specific data; for instance, it is known that the board foot–cubic foot ratio increases with diameter, and such knowledge justifies the construction of a continuously rising curve for this relationship where the data may, especially in the earlier approximations, indicate a dropping off. Such modifications will often reduce the work necessary to arrive at the final stages of the problem.

EXAMPLE OF AN ANALYSIS OF CAUSES AND EFFECT

The first illustrative example will be the problem of decomposing dextrose by sulphuric acid at high temperature.[16] This will show the type of analysis required for problems where the relationship of variables is in question. It will also illustrate the use of scale moduli in preparing the alinement chart, the influence of known facts on the shaping of the curves, and the interpretation of the results. The modification of the chart graduations will be accomplished, moreover, by a method which differs in two respects from that previously described, and which in some cases is more rapid.

OBJECTIVE, ANALYSIS, AND SOLUTION OF THE PROBLEM

The object will be to define the quantitative and qualitative relationships between the amount of dextrose decomposed by sulphuric acid at high temperatures and the factors causing the decomposition.

The various factors which influence the decomposition are: (1) The quantity of chemicals used; (2) their purity, uniformity, and stability; (3) the temperature maintained during the reaction; (4) the duration of the reaction; and (5) the uniformity of the reaction.

The necessary data can be obtained from laboratory tests. The quantities of the chemicals used in such tests can be easily determined and expressed in weight, volume or any chemical equivalent. Purity can be considered as a separate item expressed in percentages, or the quantity of the impurity may be measured. Uniformity and stability are harder to evaluate, but nonuniformity can be virtually eliminated by proper choice of materials, and the factor of stability can be ignored if all tests are conducted within a reasonably short period of

[16] This problem, while not a purely forestry problem, has been selected for illustrative purposes because it does not have the extreme complexity of most forestry problems. The problem and data were taken from Kressmann (*23, p. 30*).

time. The uniformity of the reaction can be determined by repeated tests under uniform conditions and expressed in terms of a standard error of reaction.

This investigation will be confined to the study of the effect of sulphuric acid at high temperatures upon a given quantity of dextrose exposed to its effects for a constant period of time. The quantity of dextrose and the duration of the reaction will thereby be eliminated as variable factors. Purity, uniformity, and stability of the chemicals will be eliminated as variables by using suitable materials and limiting the tests to a reasonably short period. The reaction is expected to be quite uniform and its variations, if any, are expected to be compensating. Of the factors which affect the decomposition of dextrose there remain two variables to be considered—temperature and quantity of sulphuric acid.

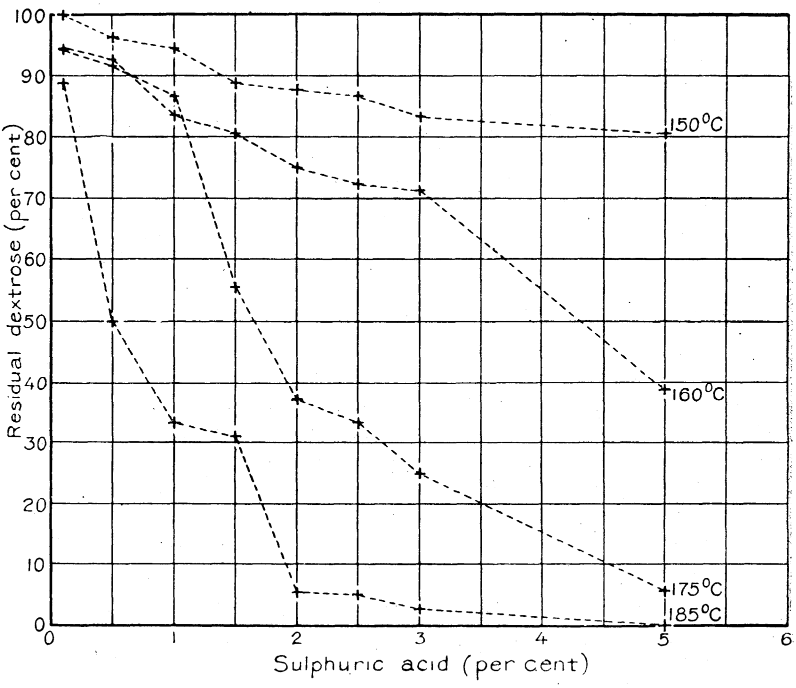

FIGURE 21.—The data of Table 8 on the decomposition of dextrose plotted in conventional form. It is apparent that the fitting of harmonized curves would be quite difficult

In addition to the factors discussed above, the results of the investigation will be influenced by experimental errors (inaccuracies in chemical analysis and instrumental and observational errors) and by personal errors in handling the data. These can not be evaluated and are therefore not susceptible to direct consideration.

Direct consideration can be given, however, to certain facts known about the problem: The amount of dextrose decomposed can vary from 0 to 100 per cent; the quantity of sulphuric acid can vary from 0 to slightly less than 100 per cent; with an acid concentration of 0 per cent no dextrose will be decomposed. All known aspects of the problem have now been enumerated and further information must be gleaned from the results of the laboratory tests.

In these tests solutions of 1 gram of dextrose in 25 cubic centimeters of sulphuric acid of various concentrations were held at various temperatures for one-half hour. The acid concentration is measured in percentages and the temperature in degrees centigrade. The amount of dextrose remaining after treatment was measured as a percentage of the original quantity. The data for 32 such tests are recorded in the first four columns of Table 8.

With but two independent variables (temperature and acid) a set of harmonized curves could be used to represent the data. In this case the preparation of such a set of curves would be difficult, as is suggested by Figure 21, even though the variables have been limited in the tests to four temperatures and eight acid concentrations. Had an irregular distribution of temperatures and acid concentrations been used it would be totally impossible to construct harmonized curves from so few data.

Beneath these data, assembled in columns 2, 3, and 4, of Table 8 (with the dependent variable dextrose last for convenience), are given the statistical measures and regression equation required, without showing the computations, which are routine in character.

TABLE 8.—*Data and computations for the dextrose problem*

Item No. (1)	(*S*) Sulphuric acid (2)	(*T*) Temperature (3)	(*D*) Residual dextrose								
			Measured (4)	Estimates							
				First (5)	Second (6)	Third (7)	Fourth (8)	Fifth (9)	Sixth (10)	Seventh (11)	Eighth (12)
	Per cent	°*C.*	*Per cent*	*Per cent*	*Per cent*	*Per cent*	*Per cent*	*Per cent*	*Per cent*	*Per cent*	*Per cent*
1	0.1	150	100.0	115.7	122.5	98.1	99.0	102.6	102.6	103.0	102.9
2	.1	160	94.4	98.3	109.1	94.9	94.5	96.6	97.0	97.2	96.7
3	.1	175	94.2	72.2	85.8	88.8	89.5	90.2	90.6	90.6	90.2
4	.1	185	88.8	54.7	61.9	76.4	81.0	80.0	79.2	79.0	79.6
5	.5	150	96.1	110.6	112.8	95.8	96.4	99.2	99.0	99.3	99.0
6	.5	160	92.7	93.2	99.3	92.4	91.8	93.1	93.2	93.3	92.8
7	.5	175	91.6	67.0	75.8	85.9	86.5	86.6	86.4	86.3	86.1
8	.5	185	50.0	49.6	51.7	59.0	67.0	64.8	62.2	61.5	63.4
9	1.0	150	94.4	104.2	101.3	92.9	93.2	95.0	94.8	95.0	94.9
10	1.0	160	83.3	86.8	88.0	89.3	88.3	88.8	88.8	88.8	88.4
11	1.0	175	86.6	60.7	64.3	79.0	79.7	78.2	78.4	78.2	78.4
12	1.0	185	33.3	43.2	40.2	33.9	40.5	40.3	36.5	35.2	38.0
13	1.5	150	88.8	97.8	91.2	90.2	90.4	91.4	91.4	91.5	91.2
14	1.5	160	80.5	80.4	77.9	86.5	84.9	84.7	84.8	84.7	84.0
15	1:5	175	55.5	54.3	54.2	64.0	63.8	61.6	62.5	61.9	62.0
16	1.5	185	31.1	36.8	30.2	15.2	18.4	20.8	19.0	17.3	20.0
17	2.0	150	87.7	91.4	82.8	87.9	87.8	88.0	88.3	88.3	88.3
18	2.0	160	75.0	74.0	69.4	82.9	78.8	77.2	78.8	78.5	78.3
19	2.0	175	37.2	47.9	45.9	46.9	44.2	43.5	45.7	44.6	45.2
20	2.0	185	5.5	30.5	22.0	7.5	8.2	10.2	10.0	8.3	10.1
21	2.5	150	86.6	85.1	76.0	85.9	85.8	85.6	86.2	86.1	85.9
22	2.5	160	72.2	67.6	62.9	77.5	69.3	67.2	71.0	70.5	70.0
23	2.5	175	33.3	41.5	39.0	31.2	28.8	30.2	34.3	33.0	33.3
24	2.5	185	5.0	24.1	15.2	3.9	4.1	5.2	5.7	3.9	5.2
25	3.0	150	83.3	78.7	70.5	83.5	83.3	82.8	84.0	83.9	83.4
26	3.0	160	71.0	61.2	57.2	69.7	58.8	56.7	61.8	61.1	60.8
27	3.0	175	25.0	35.1	33.5	20.6	19.2	21.7	25.0	23.5	24.7
28	3.0	185	2.7	17.7	9.7	1.8	1.8	2.2	3.0	1.2	2.1
29	5.0	150	80.5	53.1	59.0	72.7	74.0	71.9	75.3	75.0	75.1
30	5.0	160	38.8	35.7	46.1	47.5	36.5	37.1	42.5	41.2	40.2
31	5.0	175	5.5	9.5	22.4	7.7	8.0	10.0	12.5	10.9	11.3
32	5.0	185	0.0	−7.9	−1.5	−1.1	−.6	−2.0	−1.4	−3.4	−2.6
Total	62.4	5,360	1,970.6	1,970.7	1,976.3	1,968.4	1,952.9	1,961.4	1,989.1	1,969.4	1,978.9
Means	1.95	167.5	61.58								
SD	1.47	13.46	33.13								
SE				13.93	12.17	5.91	5.91	5.85	5.36	5.35	5.33
AC				.420							
AI					.367	.178	.178	.177	.162	.161	.161
CC				.908							
CI					.930	.984	.984	.987	.987	.987	.987

Regression equation: D= −12.7787 S −1.7426 T +378 3924.

The alienation coefficient of 0.420 indicates that only 100 − 42, or 58 per cent of the variation about the mean value of dextrose has been eliminated by using the regression equation (or the plane it represents) for estimating the values of dextrose. Figure 22 illustrates the original data and the plane representing the regression equation of—

$$D = -12.7787S - 1.7426T + 378.3924$$

The chart used is illustrated in Figure 23, B. This form of chart, with the center scale reading down and the outer scales reading up, was adopted since all variables could not be given positive signs by transposition in the regression equation.

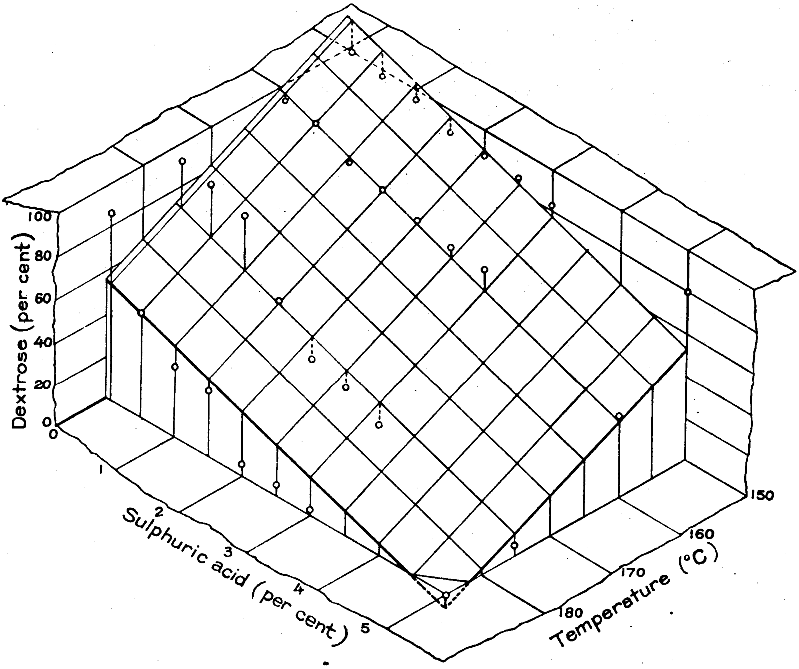

FIGURE 22.—The plane of the regression equation shown above, with the basic data indicated by circles. Note that impossible values over 100 per cent and less than 0 per cent are shown by the plane. These indicate the need of curving. Compare with Figure 25

The scale for sulphuric acid on the original chart was chosen as 0.5 per cent per inch and that for temperature as 5° per inch. The axes were located 6 inches apart. (Fig. 23 is a photographic reproduction on a reduced scale.) The axis for dextrose lies between them, and its position and scale were located by intersections as described under Construction of the Alinement Charts Needed, page 37, and checked by Formulas XI and XII.

GRADUATING CURVES

A new feature is here introduced, the graduating curve. Instead of computing and measuring the position of the dextrose graduations intermediate between zero and 100 per cent a graduation distance–graduation value curve was prepared. For such a curve the distance of each graduation from a fixed point on the axis is plotted over the value of the graduation and a curve fitted. The distances of intermediate graduations from the fixed point are read from this curve. When the interval between graduations is uniform the graduating curve is a straight line. Such is the case with this scale, and but two points are necessary to locate the line, although a third point is desirable to check the work. Figure 23, C, indicates the preparation of the graduating curve and the graduation of the *D* axis. For the dextrose axis the three points to define this line were located by intersections. Similar graduation distance–graduation value curves are prepared for each of the other axes. (Fig. 23, A and D.) These graduating curves somewhat simplify the work at this point but are chiefly useful in later stages. It will be apparent that this type of curve is merely a graphic presentation of Formula X (p. 41), to which a constant term has been added. This constant is the distance from the fixed point on the axis to the zero point of the scale.

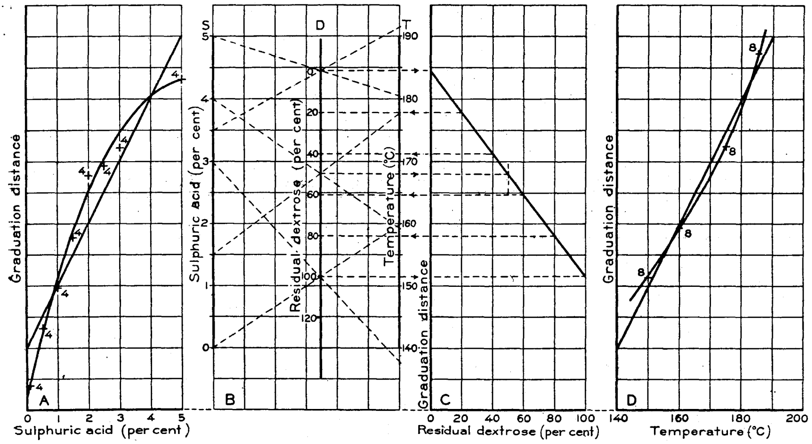

FIGURE 23.—The alinement chart for the regression equation on page 48, with the graduating curves for each axis. B is the alinement chart, with construction lines. A, C, and D are graduating curves for the three axes; the straight lines apply to the original chart while the curves in A and D result from the first estimate

Sufficient graduations should be provided on the *D* axis to permit a reading for any pair of values for *S* and *T*, even though such readings may be negative or over 100 per cent.

COMPUTATION OF RESIDUALS BY ABRIDGED METHOD

The first estimated values are now read from this chart as usual. These values are entered in column 5, Table 8. As a short cut, instead of computing individual residuals, the average residual is obtained from the totals of measured and corresponding estimated values, grouped first by acid concentration and then by temperature. The difference between the total of measured values and the total of estimated values of each class, divided by the number of items, is the deviation of the measured average from the estimated average. As usual, these deviations are considered positive if the measured exceeds the estimated, and negative if the measured is less than the estimated. Column 5 of Table 9 shows these values.

TABLE 9.—*Group-average deviations, residual dextrose*

GROUPED BY SULPHURIC ACID

Items	Class average[1]	Aggregate measured	First estimate D			Third estimate		Fifth estimate		Seventh estimate	
			Aggregate	Average residual[2]	Average residual times correction distance[3]	Aggregate	Average residual	Aggregate	Average residual	Aggregate	Average residual
(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Number	*Per cent*	*Per cent*	*Per cent*	*Per cent*	*Inches*	*Per cent*	*Per cent*	*Per cent*	*Per cent*	*Per cent*	*Per cent*
4	0. 1	377. 4	340. 9	+9. 12	−1. 46	358. 2	+4. 80	369. 4	+2. 00	369. 8	+1. 90
4	. 5	330. 4	320. 4	+2. 50	−. 40	333. 1	−. 68	343. 7	−3. 32	340. 4	−2. 50
4	1. 0	297. 6	294. 9	+. 68	−. 11	295. 1	+. 62	302. 3	−1. 18	297. 2	+. 10
4	1. 5	255. 9	269. 3	−3. 35	+. 54	255. 9	. 00	258. 5	−. 65	255. 4	+. 12
4	2. 0	205. 4	243. 8	−9. 60	+1. 54	225. 2	−4. 95	218. 9	−3. 38	219. 7	−3. 58
4	2. 5	197. 1	218. 3	−5. 30	+. 85	198. 5	−. 35	188. 2	+2. 22	193. 5	+. 90
4	3. 0	182. 0	192. 7	−2. 68	+. 43	175. 6	+1. 60	163. 4	+4. 65	169. 7	+3. 08
4	5. 0	124. 8	90. 4	+8. 60	−1. 38	126. 8	−. 50	117. 0	+1. 95	123. 7	+. 28
32	--------	1, 970. 6	1, 970. 7	--------	-------	1, 968. 4	--------	1, 961. 4	--------	1, 969. 4	------

GROUPED BY TEMPERATURE

8	150	717. 4	736. 6	−2. 40	+0. 26	707. 0	+1. 30	716. 5	+0. 11	722. 1	−0. 59
8	160	607. 9	597. 2	+1. 34	−. 15	640. 7	−4. 10	601. 4	+. 81	615. 3	−. 92
8	175	428. 9	388. 2	+5. 09	−. 56	424. 1	+. 60	422. 0	+. 86	429. 0	−. 01
8	185	216. 4	248. 7	−4. 04	+. 44	196. 6	+2. 48	221. 5	−. 64	203. 0	+1. 68
32	--------	1, 970. 6	1, 970. 7	--------	-------	1, 968. 4	--------	1, 961. 4	--------	1, 969. 4	------

[1] *S* in top section of table; *T* in lower section.

[2] The deviation of the measured aggregate from the estimate aggregate, divided by the number of items.

[3] Correction distances: For *S* scale, 0.16 inch per 1 per cent; for *T* scale, 0.11 inch per 1 per cent.

USE OF CORRECTION DISTANCES IN ALTERING THE GRADUATING CURVE

The next step is the relocation of the graduations on the *S* and *T* axes so that the dextrose readings will be greater or less than formerly by an amount equal to the departure of the measured average of each class from the estimated average. Correction distances are first determined for the *S* and *T* axes. In each case this is the distance[17] which any graduation must be moved along the axis to produce a positive change of one unit in the reading of the *D* axis. These correction distances may be determined graphically or may be computed. The general formulas for such correction distances are—

$$\text{cor. } X = \frac{XY}{ZY} L_Z \text{------------------} \text{(XIII–A)}$$

$$\text{cor. } Y = \frac{XY}{XZ} L_Z \text{------------------} \text{(XIII–B)}$$

[17] This may conveniently be in terms of graph paper divisions.

where cor. X, cor. Y denote the correction distances for X and Y, and XY, XZ, ZY are the distances between the axes and L_Z is the modulus of the Z axis. For this problem these formulas become—

$$\text{cor. } S = \frac{ST}{DT} L_D = \frac{3.0}{1.3}\left(-0.0338\right) = -0.08 \text{ inch}$$

$$\text{cor. } T = \frac{ST}{DS} L_D = \frac{3.0}{1.7}\left(-0.0338\right) = -0.06 \text{ inch}$$

L_D could be computed but was determined in this case by measuring the distance between the 0 and 100 per cent graduations and dividing by 100. By inspection, its sign is negative since the scale increases downward while those for S and T increase upward.

The negative correction distances indicate that a downward movement on the S and T axes is necessary to increase D. The algebraic products of these correction distances and the class-average deviations computed above (entered in column 6, Table 9) are measured off (according to their signs) above or below the proper graduation distance—graduation value curve, over the average temperature or average percentage of acid of the class. (The results are the plotted points in Figure 23, D and A). Smooth curves are drawn through these series of points; these curves give the revised graduation distances to be used in regraduating the S and T axes prior to making a second estimate. There are two advantages in this method. It is unnecessary to compute net regression equations or plot net regression lines, and each regression curve is drawn to such a scale (its direction may sometimes be reversed) that it may be used directly as a graduating curve. Cases may arise, however, as will be explained later, where this method is inadvisable.

The necessary graduations are next entered, lightly, with soft pencil, for they are to be erased subsequently, and the next estimate read as entered in column 6, Table 8.

The measured–second estimate curve is prepared (Table 10 and fig. 24), as in the previous example. Since it is known that the limits of dextrose are 0 and 100 per cent, this curve is so drawn that it varies only between these values. The curve through these points is now used for relocating the graduations of the D axis. The method is illustrated by the broken lines in the lower left-hand portion of Figure 24. Any reading on this axis, such as d, should be altered to read d'. This may be accomplished by placing the revised d' graduation where the d graduation was previously. These corrections may profitably be incorporated into the graduating curve (fig. 23 C) by plotting the graduation distance of d over d'. As many such points are plotted as are desired for locating this new curve of graduation distances, by means of which the new graduations are entered (in pencil, as previously done for S and T).

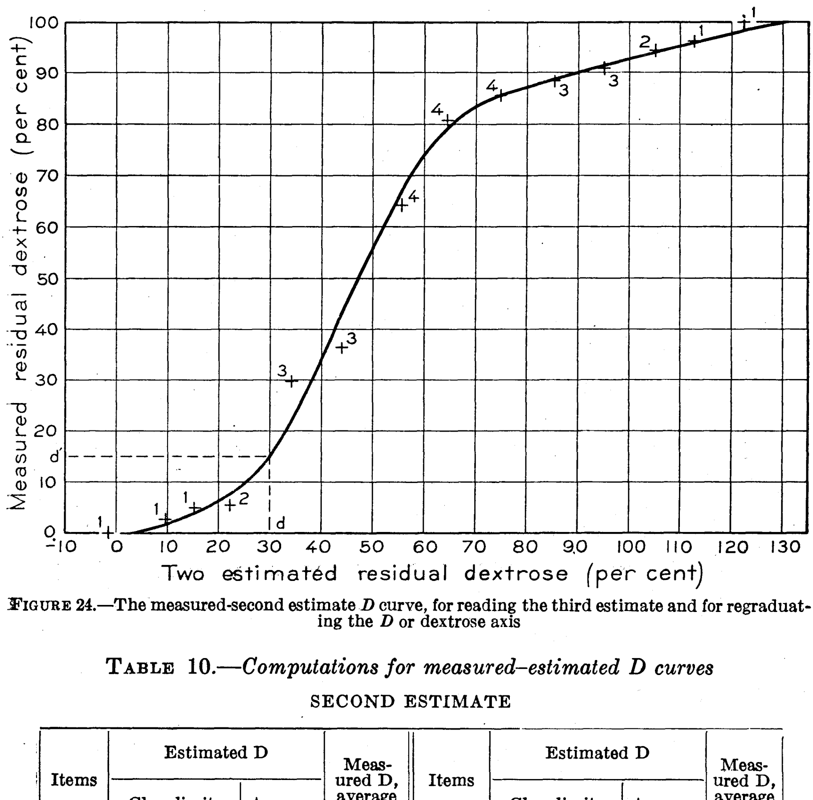

FIGURE 24.—The measured-second estimate *D* curve, for reading the third estimate and for regraduating the *D* or dextrose axis

TABLE 10.—*Computations for measured–estimated D curves*

SECOND ESTIMATE

Items	Estimated D		Measured D, average	Items	Estimated D		Measured D, average
	Class limits	Average			Class limits	Average	
Number	*Per cent*	*Per cent*	*Per cent*	*Number*	*Per cent*	*Per cent*	*Fer cent*
1	−10. 0– 0. 0	−1. 5	0. 0	4	60–0. 69. 9	64. 6	80. 6
1	. 0– 9. 9	9. 7	2. 7	4	70. 0– 79. 9	75. 0	85. 5
1	10. 0–19. 9	15. 2	5. 0	3	80. 0– 89. 9	85. 5	88. 4
2	20. 0–29. 9	22. 2	5. 5	2	90. 0– 99. 9	95. 2	90. 8
3	30. 0–39. 9	34. 2	29. 8	2	100. 0–109. 9	105. 2	94. 4
3	40. 0–49. 9	44. 1	36. 4	1	110. 0–119. 9	112. 8	96. 1
4	50. 0–59. 9	55. 5	64. 2	1	120. 0–129. 9	122. 5	100. 0

FOURTH ESTIMATE

1	−10. 0– 0. 0	−0. 6	0. 0	1	50. 0–59. 9	58. 8	71. 0
4	. 0– 9. 9	5. 5	4. 7	3	60. 0–69. 9	66. 7	59. 2
2	10. 0–19. 9	18. 8	28. 0	3	70. 0–79. 9	77. 5	80. 7
1	20. 0–29. 9	28. 8	33. 3	8	80. 0–89. 9	85. 9	87. 0
1	30. 0–39. 9	36. 5	38. 8	6	90. 0–99. 9	94. 2	94. 4
2	40. 0–49. 9	42. 4	35. 2				

SIXTH ESTIMATE

1	−10. 0– 0. 0	−1. 4	0. 0	--------	50. 0– 59. 9	----------	--------
2	. 0– 9. 9	4. 4	3. 8	3	60. 0– 69. 9	62. 2	58. 8
3	10. 0–19. 9	13. 8	14. 0	5	70. 0– 79. 9	76. 5	80. 6
1	20. 0–29. 9	25. 0	25. 0	6	80. 0– 89. 9	86. 4	85. 5
2	30. 0–39. 9	35. 4	33. 3	6	90. 0– 99. 9	94. 3	93. 4
2	40. 0–49. 9	44. 1	38. 0	1	100. 0–109. 9	102. 6	100. 0

The third estimates are read from the measured–second estimate curve of Figure 24, then grouped by temperature and acid classes, as before, and the total and class-average deviations entered in columns 7 and 8 of Table 8. The sums of the measured values, by classes, will be the same as those previously determined and need not be computed anew. Class-average deviations multiplied by the correction distances are plotted around the second graduation distance–graduation value curves,[18] which, to avoid confusion, are copied on to another sheet of paper. These curves could be plotted on the sheet on which the alinement chart is drawn, but this is apt to cause confusion, especially if many estimates or variables are necessary. The pencil graduations previously entered are erased and new ones entered (also in pencil).

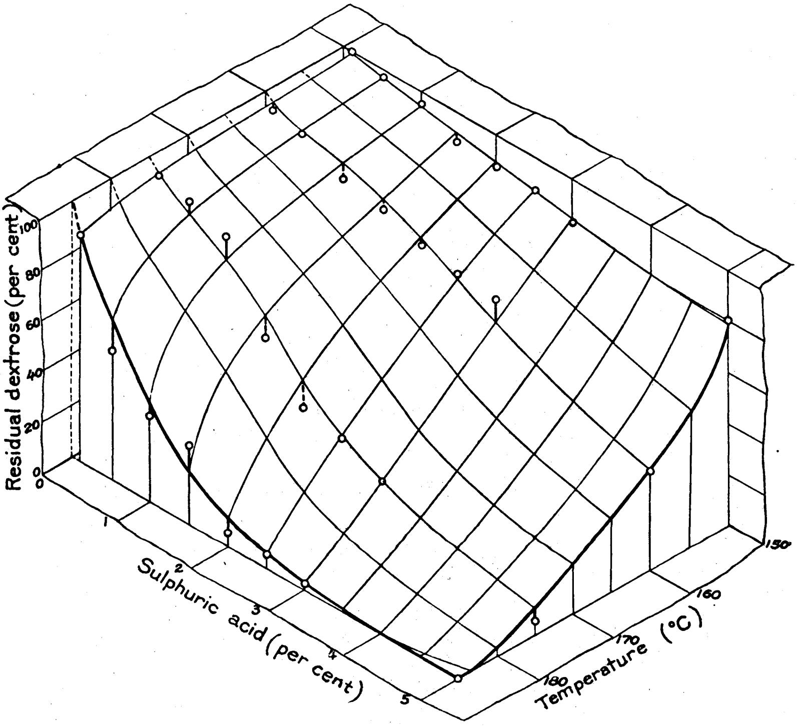

FIGURE 25.—The surface represented by the final alinement chart of the dextrose problem, with basic data shown by circles. Compare with Figure 22

The succeeding estimates present no peculiarities, and no further change is necessary after the eighth estimate. The standard error is found to be 5.33, and the final alienation index is 0.161, showing that only 16.1 per cent of the original variation about the mean value of D remains.

[18] This relocation of the independent variable graduations is not strictly comparable with the correction made after the first estimate because of the variable spacing introduced in the D axis. However, in a number of instances besides this one, this method gives better results than the more fundamentally sound but slightly less simple method described later (p. 72).

FINAL CORRECTION

The multiple alienation coefficient between the last residuals, temperature, and acid concentration, was computed, using, for the most part, the calculations made for the original regression equation. The alienation coefficient obtained was 0.9968. Were the corrections indicated made, the final alienation coefficient would be—

$$\text{final } AI = (AI_{\text{DST}})(AI_{\text{est}}) = 0.161 \times 0.9968 = 0.1605$$

indicating a reduction of variation of only 0.05 per cent, which is not sufficient to warrant modifying the chart. The final graduations are therefore entered in ink. The surface represented by the final alinement chart is shown in isometric projection in Figure 25, together with the basic data, shown by circles.

INTERPRETATION OF RESULTS

The results of this analysis show that the extent of decomposition of dextrose depends primarily upon the action of heated acid, and is not the sum of the effects of heat and acid working individually. Otherwise, the curves (fig. 25) for various values of acid or of temperature would have been similar in form. That the reaction caused by variation in one factor is not similar in extent for all values of the other factor is quite apparent from Figure 25. This conclusion would also be reached from an inspection of the final graduating curves, in which extreme reverse curvature is exhibited in that for dextrose while those for temperature and acid are somewhat similar.

The remaining variation in the data, 16.1 per cent of the original variation, may be attributed, among other things, to discrepancies in measurements of materials and in analysis of chemicals, to variations in temperature during the experiment, and to nonuniformity of materials.

EXAMPLE INVOLVING THE DEVELOPMENT OF A PREDICTING MECHANISM

An example of the second type of problem, in which a predicting mechanism is the primary object, is a study of bark thickness. The object of the study was to find the factors giving the best indication of bark thickness (of shortleaf pine) and to incorporate them into a mechanism for predicting bark thickness at any point on the stem.

The analysis of the problem and the method of selecting the measurements to be employed will be given. The construction of a chart for more than three variables, employing one axis twice, and the combined use of regression lines and graduating curves will be new features illustrated. The effect of high intercorrelation between independent variables will be shown, and the approximate standard error will be used for measuring improvement in predictions .

The factors affecting bark thickness are of two types, those causing bark growth and those causing removal of its outer surface. Direct measurement of either type of factor is difficult, and for a predicting mechanism recourse must be had to indirect measurements. Of the causes of cambium activity resulting in bark growth the most usable and easily obtained measurements are age, site index, and dimensions of the tree. Of the causes of bark removal—such as furrowing and action of frost, wind, rain, and fungi—no good indicator can be obtained. Age and site index are the best, but site index is chiefly a measure of combined soil and climatic effects during the growing season, whereas bark reduction is more the effect of climate alone throughout the entire year. However, site index will be given consideration in the initial steps.

A list of the measurements from which it may be possible to predict bark thickness[19] will include site index, age, diameter at breast height, total height of tree, and diameter of the stem at various heights (both absolute and as a percentage of total height). Several other factors might be listed, such as crown class, crown vigor, crown density, locality, aspect, and heredity, some of which are incommensurable. Crown class, vigor, and density may be evaluated by some expression of current annual growth (rings per inch, current annual increment, etc.), but this does not measure their cumulated effect upon bark growth, whereas diameter more nearly does. Locality and heredity do not lend themselves to easy evaluation, while aspect, in terms of azimuths, is a very unsatisfactory value to work with. These six last-named variables hold little promise and would complicate the predicting mechanism. They are accordingly rejected. To determine which of the remaining variables have greatest value for predicting purposes a multiple regression equation (A), including all the remaining variables, was computed with bark thickness as the dependent variable. The equation obtained was—

A)—Bark thickness (inch)	=0. 000775 site index	+0. 0626 d. b. h.	−0. 00752 total height.
Range of variable =	54 feet.	9 inches.	72 feet.
Range×coefficient =	0. 0418 inch.	0. 563 inch.	0. 541 inch.
	+0. 00519 total age	−0. 0207 section d. i. b.	−0. 000941 section height (feet).
	68 years.	15 inches.	85 feet.
	0. 333 inch.	0. 310 inch.	0. 0800 inch.
	−0. 00987 section height (per cent)	−0. 00447 section age	+0. 751.
	100 per cent.	80 years.	
	0. 987 inch.	0. 358 inch.	

The multiple alienation coefficient was 0.575.

The range of measurements (difference between maximum and minimum) is given below the equation, and below that is given the corresponding variation in bark thickness (range of measurements times regression coefficient). Considering both the regression coefficient and the maximum variation in bark thickness associated with each variable, the use of site index, tree and section ages, and section height in feet will not increase the accuracy enough to offset the increased number of measurements needed nor the increased complexity of the predicting mechanism.[20] Accordingly these variables are rejected and a new regression equation (B) computed for those remaining.[21]

This equation was—

(B)—Bark thickness (inch)

$$=0.070339 \text{ d. b. h} - 0.006002 \text{ total height} - 0.033235 \text{ section d.i. b.} - 0.009778 \text{ section height (per cent)} + 0.78716.$$

[19] The correlation of diameter outside bark with diameter inside bark (and other variables) is an alternative treatment. This was not used, however, because the diameter inside bark is the major component of diameter outside bark, and high correlation naturally would be obtained. Significant differences in bark thickness would appear as very minor variations in the relation between diameter outside bark and inside bark, thus obscuring the presence of any factor having an appreciable influence on bark growth.

[20] Section d. i. b., while apparently less important than total age and section age, must be retained for practical reasons. It will be needed if diameter outside bark at points up the tree are for any reason to be calculated.

[21] The squares, products, sums, and standard deviations necessary have already been computed in connection with the preceding equation, so relatively little work is necessary for the new equation.

The alienation coefficient was 0.577, only 0.002 more than for all variables, a fact which confirms our judgment that four variables may be dropped without appreciable loss in accuracy. Although the regression coefficient for total height is now small it is not certain that a curvilinear relation does not exist, and it is therefore retained.[22]

Of the four independent variables retained, it is quite likely that one of the sectional measurements (section d. i. b. or percentage height) might be reflected in the other should either be removed from consideration. With total height and d. b. h. considered there should be moderately high correlation between section d. i. b. and percentage height, although form varies between trees of the same height and d. b. h. To analyze this two regression equations were computed with section d. i. b. and section height (per cent) omitted in turn. These were—

(C)—Bark thickness = 0.0498 d. b. h. − 0.00666 total height − 0.00734 section height (per cent) + 0.702

(D)—Bark thickness = 0.00366 d. b. h. − 0.00862 total height + 0.0773 section d. i. b. + 0.434.

The alienation coefficient was 0.590 for equation C and 0.724 for equation D, as compared with the 0.577 for equation B, with both section d. i. b. and section height (per cent) included. The small difference between the coefficients for equations B and C indicates that in applying the result to trees of average form the three measurements d. b. h., height, and section height (per cent) are sufficient. If form class were recognized, it would be necessary to retain both section d. i. b. and section height (per cent) or to include form quotient in equation C. Obviously, a predicting mechanism based on equation D, with an alienation coefficient of 0.724, would be less satisfactory than one based on equation B or C.

This particular study was for application to trees of average form, but equation (B) including both section d. i. b. and section height (per cent) will be developed because it illustrates several points not encountered in the development of equation (C), the more satisfactory of the 4-variable equations.

THE ALINEMENT CHART

An alinement chart for five variables requires seven axes,[23] one for each independent variable, one each for the two intermediate sums, and one for the final sum (the dependent variable). By proper design some of these axes may be made to coincide, thus reducing the number appearing on the chart.

The initial chart (fig. 26) was constructed as follows, to agree with the regression equation B:

The chart was assembled progressively, starting with two of the independent variables and adding one variable at a time. Because of the possibility of magnified and cumulative errors due to slight inaccuracies in mechanical drawing, the graphic method of locating

[22] The variables rejected above could have been tested for curvilinearity by carrying them through one or two approximations, but the large number of data together with the small likelihood of developing appreciably curved regression lines led to the omission of such a test in this case. Also, site index and ages are closely associated with the retained measurements of the tree, and are thus included by implication. Height of the section, in feet, is also included in another form in the two measurements of total height and percentage height.

[23] The number of axes required is 3 less than twice the number of variables.

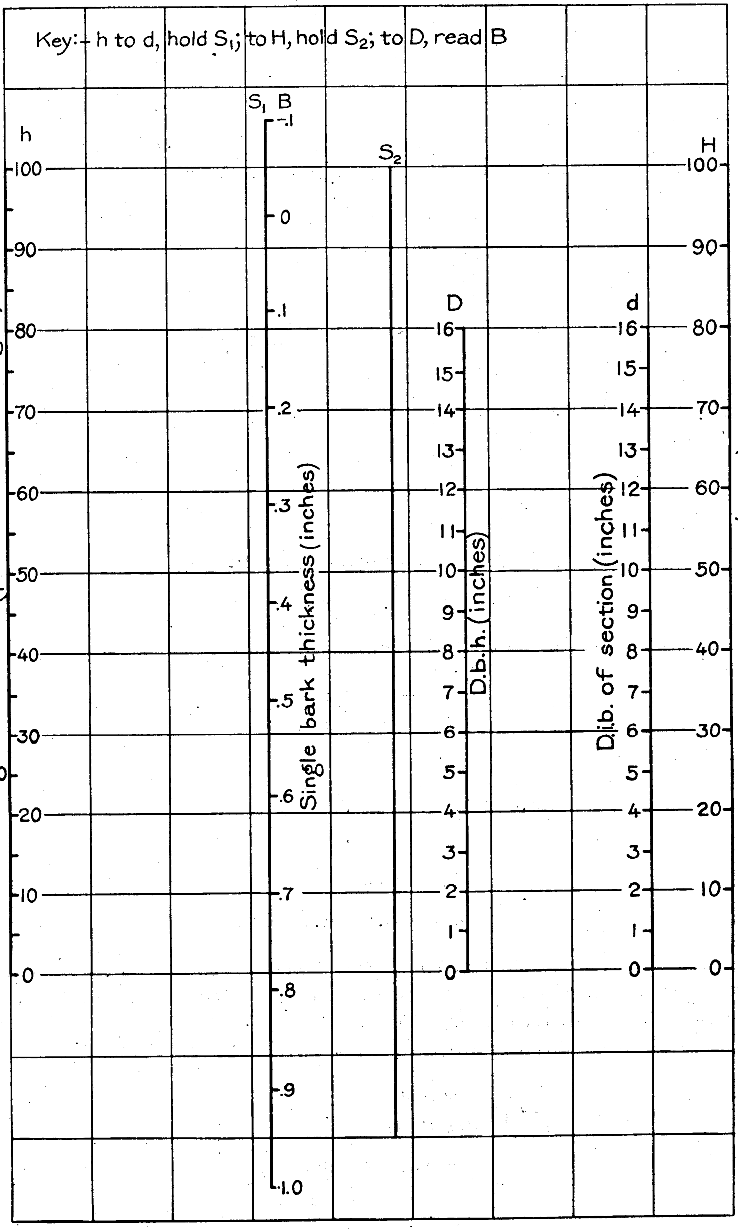

FIGURE 26.—The alinement chart for the bark-thickness regression equation B:

$$B = 0.070339D - 0.006002H - 0.033235d - 0.009778h + 0.78716$$

The *B* axis is also used to hold S_1 values, which need not be read and hence require no scale the position of the axes was abandoned in favor of calculated values. For the start two variables alike in sign (a matter of personal preference) were selected, namely, section height (per cent) as *h*, and section d. i. b. as *d*, both with negative coefficients. The sum axis (S_1) for these two components of the chart will be between the others if all scales increase in the same direction. For convenience, these scales will be made to increase from the bottom to the top. The distance between *h* and *d* was made 8 inches, the scale for *h* was made +0.1 inch for each percentage unit, and for *d*, +0.5 inch per inch of diameter. From Formula X (page 41) we have—

$$U_h = L_h f(h)$$

$$L_h = \frac{U_h}{f(h)}$$

$$L_h = \frac{0.1h}{(-0.009778h)} = -10.23$$

In a similar manner—

$$U_d = L_d f(d)$$

$$L_d = \frac{0.5d}{(-0.033235d)} = -15.04$$

Since the distance *hd* has been made 8.0 inches—

$$hS_1 = hd - S_1d = 8.0 - S_1d$$

Now, from Formula XII (p.42)—

$$\frac{hS_1}{S_1d} = \frac{L_h}{L_d}$$

Substituting—

$$\frac{8.0 - S_1d}{S_1d} = \frac{-10.23}{-15.04} = 0.6802$$

$$S_1d = 4.76$$

$$hS_1 = 8.0 - 4.76 = 3.24$$

The S_1 axis is accordingly entered parallel to and 3.24 inches to the right of the *h* axis. Its scale need not be entered since no values are to be read from it, but the modulus will be needed for subsequent calculations, and so is computed as follows:

$$L_{S_1} = \frac{L_h L_d}{L_h + L_d}$$

$$= \frac{-10.23(-15.04)}{-10.23 + (-15.04)}$$

$$= \frac{153.86}{-25.27} = -6.09$$

To the chart as it now stands an additional axis (*H*) for another independent variable is added. This is done by considering the

S_1 axis as one component of a new chart and H as the second component. Another axis (S_2) will be needed for their sum. Total height, with a negative coefficient −0.006002, will be taken for H. The axis for H will be arbitrarily placed 5.76 inches to the right of S_1, with a scale of +0.1 inch per foot, increasing in the same direction as the preceding ones. S_2 will therefore lie between S_1 and H. The following computations give its position and modulus—

$$L_H=\frac{U_H}{f(H)}=\frac{0.1H}{(-0.006002H)}=-16.66$$

$$\frac{S_1S_2}{S_2H}=\frac{S_1H-S_2H}{S_2H}=\frac{5.76-S_2H}{S_2H}=\frac{L_{S_1}}{L_H}=\frac{-6.09}{-16.66}$$

$$S_2H=4.22$$

$$L_{S_2}=\frac{L_{S_1}L_H}{L_{S_1}+L_H}=\frac{-6.08(-16.66)}{-6.08+(-16.66)}=\frac{101.29}{-22.74}=-4.45$$

There now remains one independent variable (d. b. h.) to be added to this chart. S_2 is then considered as the first component of a new chart, d. b. h. (D) as the second component, and their sum will equal bark thickness (B) minus the constant (by transposition in the regression equation.) Since the regression coefficient for D is positive it would be necessary, in order to have all scales increase in the same direction, to transpose the equation as follows—

$$-S_2+D=(B-\text{constant})$$

$$-S_2-(\text{B}-\text{constant})=-D$$

which then places the D axis between those for S_2 and B. However, it is desired to make the B axis coincide with the S_1 axis, and there is rather too little space for D between S_1 and S_2 ($S_1S_2=S_1H-S_2H=5.76-4.22=1.54$ inches). Accordingly, the B scale is made to read in the opposite direction from the others. By reversing the sign of B (and hence the direction of the scale) the equation becomes—

$$-S_2+D=-(\text{B}-\text{constant})$$

$$S_2=D+(B-\text{constant})$$

thus making S_2 the central scale. The scale for D was taken as +0.5 in. per inch of d. b. h. Since the moduli of S_2 and D have been fixed, as have the positions of the B and S_2 axes, Formula XII is used for finding BD.

$$L_D=\frac{0.5D}{0.070339D}=7.11$$

$$\frac{S_2B}{BD}=\frac{L_{S_2}}{L_D}=\frac{-4.45}{7.11}=-0.626$$

Since $S_2B=S_2S_1=-S_1S_2=-1.54$

$$\frac{-1.54}{BD}=-0.626$$

$$BD=2.46$$

D is therefore placed 2.46 inches to the right of *B*. The modulus of *B* is—

$$L_B=\frac{L_{S_2}L_D}{L_{S_2}+L_D}=\frac{-4.45\times7.11}{-4.45+7.11}=\frac{-31.64}{2.66}=-11.9$$

The key to the chart is, then,

From *h* to *d*, hold S_1 (*B*); to *H*, hold S_2, to *D*, read *B* (S_1).

To graduate the *B* axis, the elevation of any graduation of *B* should be determined and plotted over its value and a line drawn through this point with a slope of −11.9 inches per inch of bark. From this graduating curve the scale of *B* is entered, thus completing the chart.

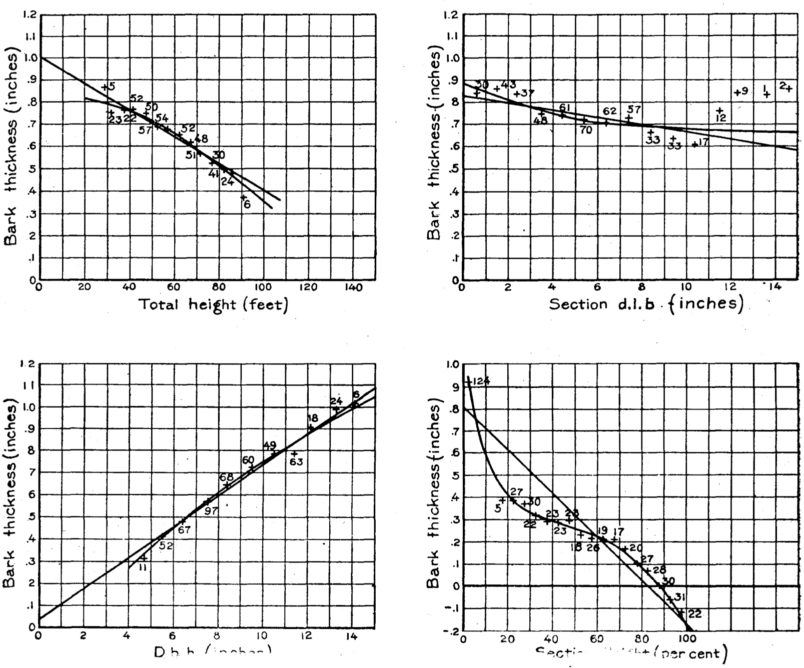

FIGURE 27.—The regression straight lines, first estimate residuals and regression curves for bark thickness

First estimates for all items are next made, the measured and first estimated values are grouped by independent variable classes, and the class-average deviations are computed. Correction distances, however, will not be used in this problem. This is advisable because with many-variable charts, the correction distance for some of the variables may become too large, and the dispersion of residuals so plotted may be too great to permit of the fitting of a curve with facility. In such cases the regression-line method used in the first illustrative problem (p. 24) should be reverted to. In this study regression lines will be used for all variables, although the correction distances for *H* and *D* are still small enough to permit of their use.

REGRESSION LINES

The net regression straight lines are drawn (see p. 25), deviations plotted about them, and regression curves fitted. (Fig. 27.) It will be noted that those indicated for *D* and for *H* curve in the same direction; and since d. b. h. and height are known to be intercorrelated it is well not to put in all of the curvature indicated, lest some may have to be removed in later approximations. The curvature for *h* is considerable, but well defined. That for *d* tends to rise toward the right, after dropping from the left. This right end is rather poorly defined, and it is quite probable that the rise indicated is accidental. Accordingly, this rise will not be put in until its correctness is confirmed by subsequent estimates. The *d* curve is, therefore, kept dropping toward the right, instead of rising as the points indicate.

GRADUATING CURVES FROM REGRESSION CURVES

The graduating curves corresponding to these regression curves are next prepared by locating a series of points as follows (fig. 28):

For a section height of 10 per cent the value of bark thickness is read from the new regression curve. The section height for which the original regression line gives the same bark thickness is next determined to be 23 per cent. On the graduation distance–graduation value curve for *h* a point is plotted over 10 per cent at the same height as the original 23 per cent. A number of points are thus located and the new graduating curve drawn through them. Graduating curves for each independent variable are thus prepared, the alinement chart regraduated and a second estimate read.

Its approximate standard error[24] was 0.117, which is to be compared with the standard deviation of 0.300 and with 0.169, the true standard error of the first estimate. As in the previous examples a measured–

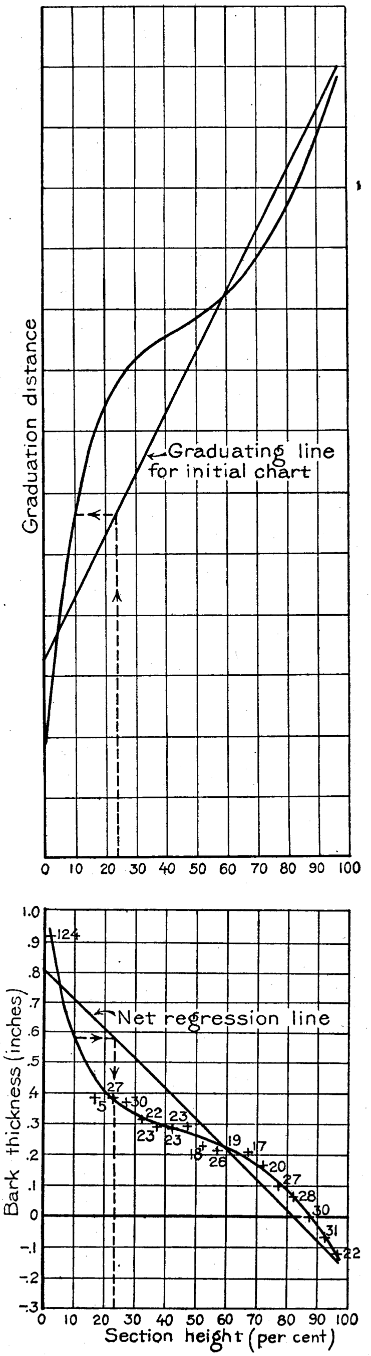

FIGURE 28.—Revising the graduating curve, for section height (*h*), from the regression curve. The method is illustrated in the location of the 10 per cent point

[24] Because of the large number of data, the improvement with each estimate was indicated by the approximate standard error, 1.253*AE*. *AE* can be obtained as shown under Short Cuts in the Appendix without computing the deviation of each item. Its use with small numbers of data is inadvisable, since it may fail to eflect minor improvements.

second estimate curve is prepared, the bark-thickness axis regraduated and a third estimate read. The third estimate is treated as was the first, the deviations being plotted (fig. 29) about copies of the second regression curves (broken lines) to avoid confusion.[25] The D and H curves (not shown) each indicate increased curvature, which is entered. The h curve retains its general outline, but reduces its curvature slightly. The curve for d reverses the direction of curvature, indicating that it is closely correlated with another variable, h.

EFFECT OF INTERCORRELATION OF TWO INDEPENDENT VARIABLES

The subsequent curves for D and H show less change in curvature and have become quite stable in position. Those for d and h give evidence of an oscillation in curvature as shown in Figures 29 and 30. If the change in curvature indicated is only partly entered, they will eventually reach a stable position. However, even though these two curves continue to shift, the shift is compensating in effect, and the accuracy of the chart is not affected, as is shown by the standard errors, which, for the third, fourth, and fifth estimates (0.124, 0.123, and 0.121, respectively), are almost identical, although these two regression curves have shifted considerably.

At this stage a multiple correlation of the residuals should show whether or not a tilt of the regression curves would eliminate the swinging. Such a correlation was computed, resulting in an alienation coefficient of 0.955 and a regression equation of—

$$\text{Bark thickness residuals (inch)} = 0.000400D - 0.0000862H - 0.00162d - 0.000132h + 0.0174$$

The tilt indicated by the coefficients is very small for all variables. The final alienation coefficient, were the indicated corrections made, would lower the alienation coefficient from 0.403 to—

$$0.403 \times 0.955 = 0.385,$$

an improvement too small to warrant the effort necessary. The corresponding improvement in the standard error would be from 0.121 to 0.116.

A predicting mechanism has been developed, enabling the calculation, with satisfactory accuracy, of bark thickness and hence of outside-bark diameters, for all points on the stems of second-growth shortleaf pine of average form. Inside-bark stem measurements must be known, of course, to apply the results when bark volume or outside bark diameters are desired.

A simpler chart, easier to make and to use, could doubtless have been developed with the variable of section diameter omitted, thus eliminating the oscillation of the regression lines. For some purposes such a chart would be adequate, but for computing bark volume or outside bark diameters additional information as to stem measurements is as essential for its application as with the chart which has been prepared. It might also be advisable—were the object the development of the best predicting mechanism rather than the illustration of a method—to investigate the possibilities of an even simpler chart in which but two independent variables, D and h, are used.

[25] The original straight regression line is copied also, for use in regraduating the chart, since it is some what easier to refer the new curve to the straight line.

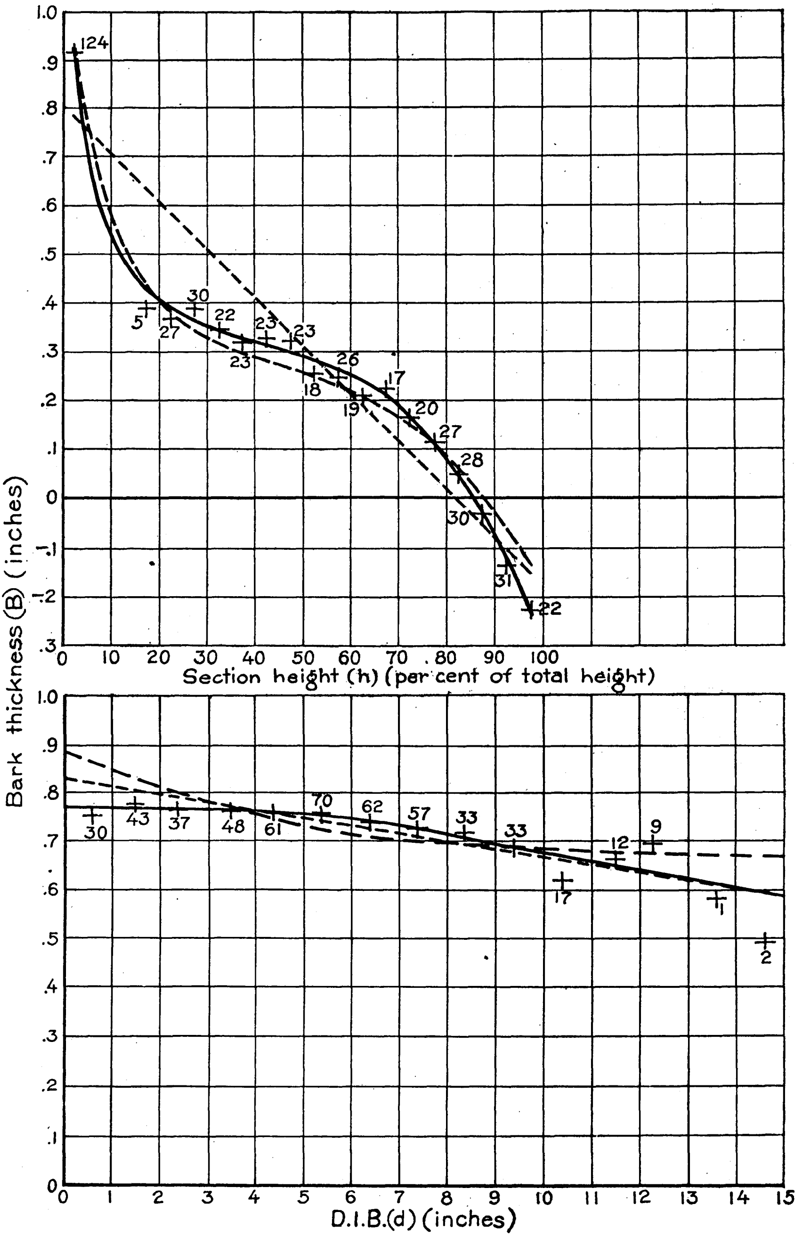

FIGURE 29.—Residuals of the third estimate plotted about copies of the second *d* and *h* regression lines. The first (straight) regression lines are included to aid in regraduating the chart. Note the changes in curvature, which do not persist in the following steps

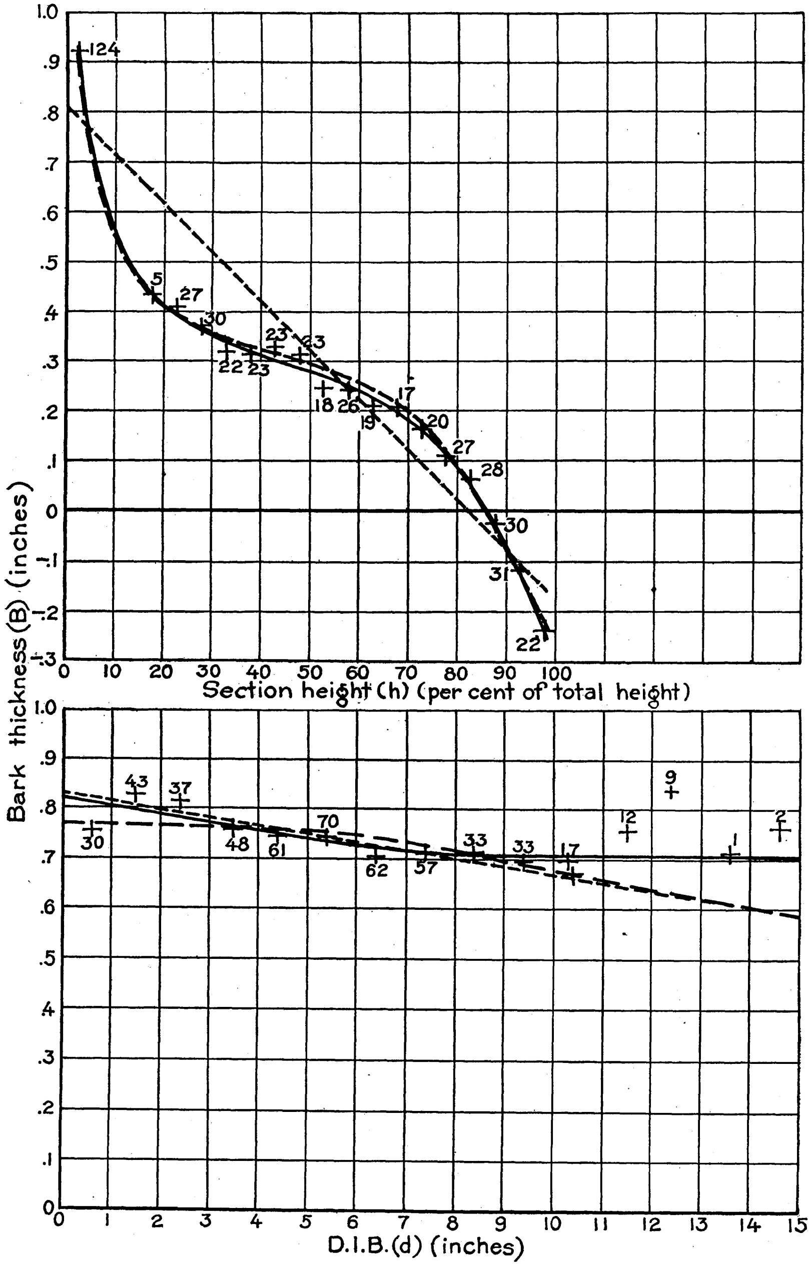

FIGURE 30.—Section diameter (*d*) and percentage height (*h*) regression curves resulting from the fifth estimate. A comparison with Figure 29 shows an oscillation of curvature. This, caused by a high intercorrelation between these variables, may be checked by inserting only a part of the indicated curvature in each case

COMPLEX PROBLEMS AND INADEQUATE DATA

The difficulty of interpreting analyses of complex problems, particularly when based on records in a form not suited to the method of analysis, will be illustrated by a study of the damping-off of coniferous seedlings[26] in which the variables have a periodic character.

The data obtained comprised twice-daily records (1) of the percentage of seedlings showing symptoms of damping-off, and continuous records of (2) evaporation, (3) soil temperature (curved), (4) air temperatures, and (5) soil moisture.

The objective of the problem was the definition of the type of variation in damping-off associated with each of the above factors or causes contributing to the appearance of damping-off symptoms. Air temperature was not used because it is reflected in the evaporation data. Variables will, therefore, be restricted to (1) number of days since germination of first seedling, (2) evaporation, (3) soil moisture, (4) soil temperature, and (5) number of damped-off seedlings, as the dependent variable.

Of these variables, two are expressed as rates; number of seedlings damped-off is given as the percentage (of initial number) damped-off per hour; evaporation is given in cubic centimeters per hour; soil moisture as a percentage; soil temperature in degrees Fahrenheit.

There is a cycle 24 hours in length involved in each of the last four variables, but only for the last three variables have the data been taken in such form that the cyclic fluctuations can be analyzed. In recording the data for damping-off in terms of average values for day and night this cyclic fluctuation is obscured and it therefore becomes necessary to treat the other variables in a comparable manner to avoid erroneous results. Accordingly, average values for each period are used for soil temperature, soil moisture, and evaporation. This procedure is dictated by the character of the records and. does not represent the most desirable method of analyzing this type of problem.

The regression equation computed for these data was—

Number of seedlings damped-off (percentage of initial number) = $-0.0469 \times$ age (days) $- 0.0069 \times$ soil temperature (° F.) $+ 0.0288 \times$ evaporation (cubic centimeters per hour) $- 0.0039 \times$ soil moisture (per cent) $+ 1.4713$

The alienation coefficient was 0.785.

Proceeding with the analysis, in the way already described, the regression curves shown in Figure 31 were obtained with the sixth estimate. The final alienation coefficient was 0.403.

The curves for age, evaporation, and soil temperature appear to be reasonable; the older seedlings are more hardy, and fewer of the diseased seedlings will show decided symptoms, their increased stiffness making detection of the diseased condition harder. Increased evaporation increases transpiration and the wilting indicative of damping-off becomes more apparent. The increase in disease-progress with an increase in temperature, up to about 70° F., is in accordance with general knowledge of fungus growth. The first curve, for soil moisture, is rather peculiar, since it is known that the

[26] The data for this study were made available through the courtesy of Carl Hartley (*14, p. 75–79*), of the Bureau of Plant Industry.

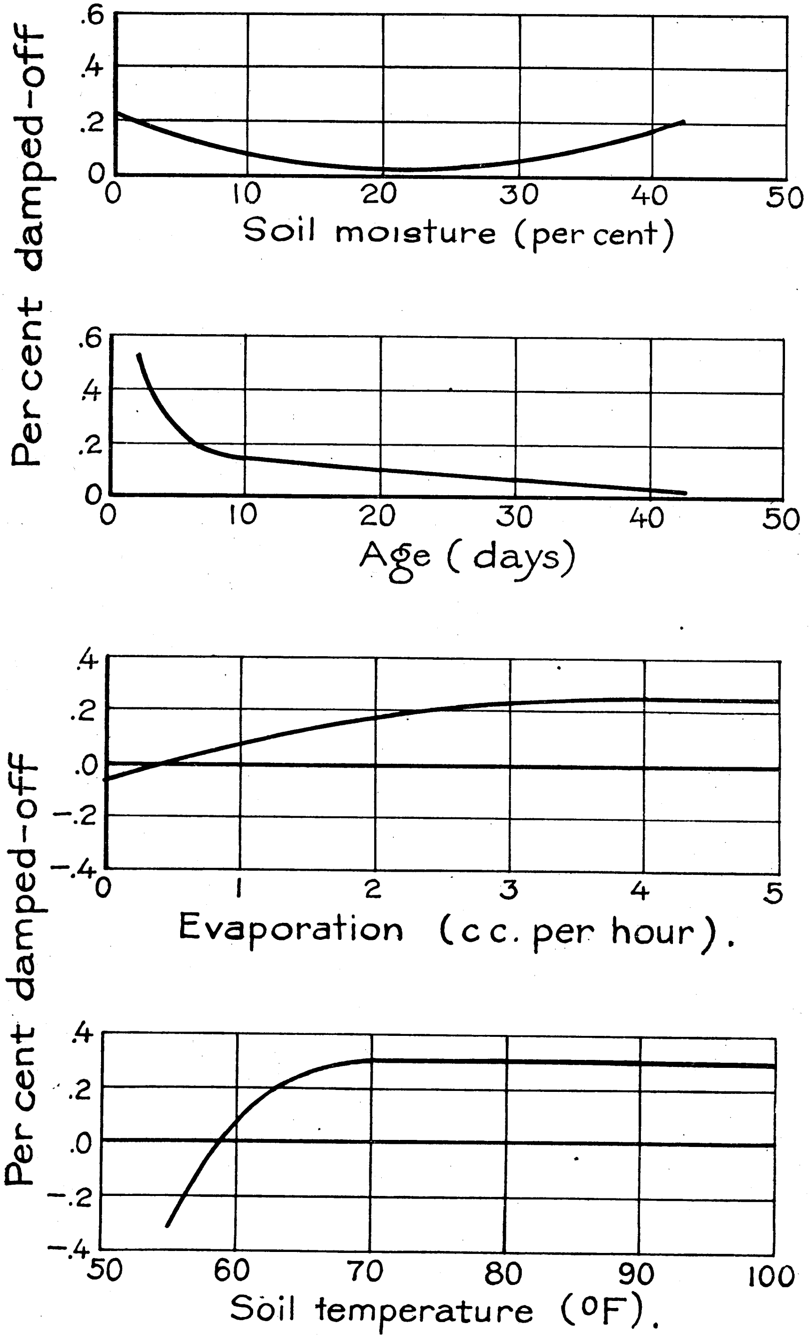

FIGURE 31.—Final regression lines (from sixth estimate) for the damping-off problem

disease progresses more rapidly in wet soil than in dry. Several interpretations may be given. Since the dependent variable is the number of seedlings showing symptoms of damping-off (wilting), one interpretation may be that when soil moisture is near the wilting coefficient a number of nondiseased seedlings may wilt, thus indicating a larger number affected when the soil moisture is low. The portion of the curve above 20 per cent is in accord with the known relation of moisture to the disease growth.

The shape of this curve lends little support to a suggestion made by Hartley, in his discussion of the data, that increased soil moisture will reduce the amount of air in the soil, thus reducing the disease activity. This decrease of air in the soil is a variable which has been considered only by implication, through its association with soil moisture. This hypothesis is reasonable, but seems to be contradicted by the rising right-hand portion of the curve.

Several disturbing factors may be responsible for the peculiarity of this curve. One is associated with its data; the moisture determinations were limited to the top 0.6 centimeter of soil and may not be entirely indicative of the moisture content throughout the entire zone in which the disease was active. The second disturbing factor is that the present value of the dependent variable may depend on previous values of the independent variables as well as on their present values. The symptoms of damping-off may not be related entirely to the soil moisture during any particular instant or short period, but may rather be related to the moisture content at some previous period, or to the accumulated effects during a longer period, or to an optimum or critical value, and to the time or to the average value since the optimum or critical point was reached.

It is evident, therefore, that a very complex relation may exist, not only in connection with soil moisture, but with evaporation and temperature as well. A certain optimum combination of these three variables may produce a maximum effect at a later date, depending upon the conditions obtaining during the interval. The data at hand are obviously inadequate, and additional information is necessary as to soil temperature and moisture at other depths, as well as a more continuous record of damping-off in which bona fide damping-off is distinguished from ordinary wilting.

TIME SERIES

The correlation of time series, in which values of one variable at given dates are correlated with values of other variables at those dates, is a most difficult problem for statistics. Experience, care, and extreme caution are necessary to avoid nonsense correlations (*48*). Such problems often involve periodic variations, which have commonly been eliminated before correlating, as in the problem on damping-off (p. 65). Errors are often introduced by such procedure and should be studiously avoided. The use of serial correlations (*41*) and, in curvilinear correlation, the use of seasonal and secular variations as additional variables offer a safeguard against illogical results. This phase of statistics is so complex and undeveloped that its presentation in this bulletin is inadvisable. The difficulties affect all graphical and statistical methods and are not peculiar to the technic herein described.

For the solution of such complex problems it must suffice to point out the possibility of using maxima (*25*, *26*), minima, modes and means, seasonal trend and secular variation together, moving averages, accumulations, values at previous times (lagging data), interval since a given occurrence, such as rainfall.

MINOR VARIATIONS IN TECHNIC

USE OF RATIOS

The basic data for many problems may often be expressed both in absolute quantities and in ratios or percentages. Ratios, or percentages, may have two uses, (1) as a more logical measure of the variable, as in the dextrose problem where the relationships change with the proportion, rather than the quantity of each chemical; and (2) to simplify the problem by reducing magnitude or range of numbers, or by reducing curvature of the regression lines. In the first instance the ratios themselves are the data being investigated, while in the second the ratios serve merely as a means to an end.

The second use occurs chiefly in constructing charts for predicting purposes. To consider such a chart satisfactory it is essential that, when checked against its basic data, the sum of the values estimated by means of the ratios shall not be appreciably higher or lower than the sum of the measured values in absolute units. Because of an uneven distribution of data a difference between these sums may exist even when the sums of the estimated and measured ratios are exactly equal. A correction of the estimates becomes desirable in such instances and may be made in one of two ways. The simplest way is to multiply each estimate by the ratio $\frac{\text{Sum measurements (absolute units)}}{\text{Sum estimates (absolute units)}}$. The second type of correction, more desirable when numerous estimates are made, is a modification of the scale for the dependent variable, to increase or decrease the readings of the chart by the ratio given above. Thus, if the ratio were 1.02, the 1.00 graduation becomes 1.02, the 10.00 graduation becomes 10.20, the 50.00 graduation becomes $50.00 \times 1.02 = 51.00$, etc. A graduating curve through a few such points will permit a regraduation of the entire scale.

GROUPED DATA

The basic data for some problems may be so numerous that a prohibitive amount of work would be involved in making a separate estimate for each item. Averages by classes, such as diameter-height, age-site index classes, may be used in such cases, but all residuals must be weighted (multiplied) by the number of items in each class. When this is done a standard error of the class averages may be computed for following the improvement in estimates, but such a standard error is meaningless as an expression of the variation in the original data, since the deviation of the averages is influenced greatly by the number of items entering into each average. For a true standard error a final estimate should be made for each item and the standard error computed from their deviations. When many variables are involved the number of items in each class may be so small that the time involved in weighting the averages overbalances the savings in number of readings of the chart.

INCOMMENSURABLE VARIABLES

If some factors are not commensurable they may be ignored in the initial steps and their effects later analyzed as follows. A composite chart is prepared for all data, using commensurable variables only. The residuals of the final estimate from this chart are grouped by the recognized classes of the incommensurable variables (crown classes, localities, species, collector of the data, etc.). The departure of the average residual for each of these classes is a measure of the effect of the variable.

EFFECT OF HIGH INTERCORRELATION OF INDEPENDENT VARIABLES

Occasionally the existence of very high intercorrelation between two of the independent variables will cause serious difficulties. This can best be illustrated by an extreme case, based on hypothetical data in which two of the variables, X and Y, are by intent more closely correlated than they would be in any actual forestry problem. A group of values was calculated by means of the formula—

$$\left(\frac{W}{2}\right)^2 = -\sqrt{X} - 10 \log Y + Z^2$$

A regression equation for these values was computed to be—

$$W = +0.466X - 0.958Y + 3.82Z - 4.67$$

The alienation coefficient was 0.290 and the standard error 1.65. This regression equation has the sign for X opposite to that in the basic equation, showing that it is seriously in error.

After modification for curvature the alienation coefficient was lowered only to 0.276, from 0.290, and the standard error was 1.57, as compared with 1.67, the standard error of the first estimate. The discrepancy in the sign of X remained.

This failure of the method is less serious than might first appear if a predicting mechanism is the chief aim. If one of the two variables which are highly intercorrelated is omitted, its influence will be carried into the final estimates through that which remains, its final regression curve being essentially a composite of the direct effect of one and the indirect effect of the other. In such cases a predicting mechanism is produced which, while based on but one of these variables, is nearly as accurate as if both highly intercorrelated variables had been retained.

VARIABLES WITH SMALL REGRESSION COEFFICIENTS, BUT WITH CONSIDERABLE CURVATURE

In certain problems it may sometimes be desirable to retain a variable when its linear regression line has little slope, but considerable curvature is expected, for in such cases the final influence of the variable may be considerably greater than is indicated by its regression coefficient. In preparing the alinement chart, a very small scale unit should be used for the axis representing such a variable, for subsequent corrections will almost certainly expand the scale, and sometimes to an astonishing degree. A relatively small scale unit on one axis can be secured by locating it very close to the sum axis associated with that variable. In such problems. moreover, the use of correction distances, as described in previous pages, usually results in a wide scattering of the points which makes curve fitting difficult. The method used in the bark-thickness problem—plotting the residuals around the net regression line, curving, and then transferring this curvature to the graduating curve—should therefore be substituted.

The nearly horizontal position of the regression line may make an accurate transfer of the regression curve to the graduating curve difficult. To overcome this difficulty it may be desirable to copy both regression line and curve on to a new graph in which a much larger vertical (and smaller horizontal) scale is used, thus steepening the lines and facilitating the reading of their intersections with the with the horizontals of the cross-section paper.

ASSUMED CHARTS

The initial chart for each of the problems presented in these pages has been based on a regression equation, and is often radically changed in the subsequent steps.

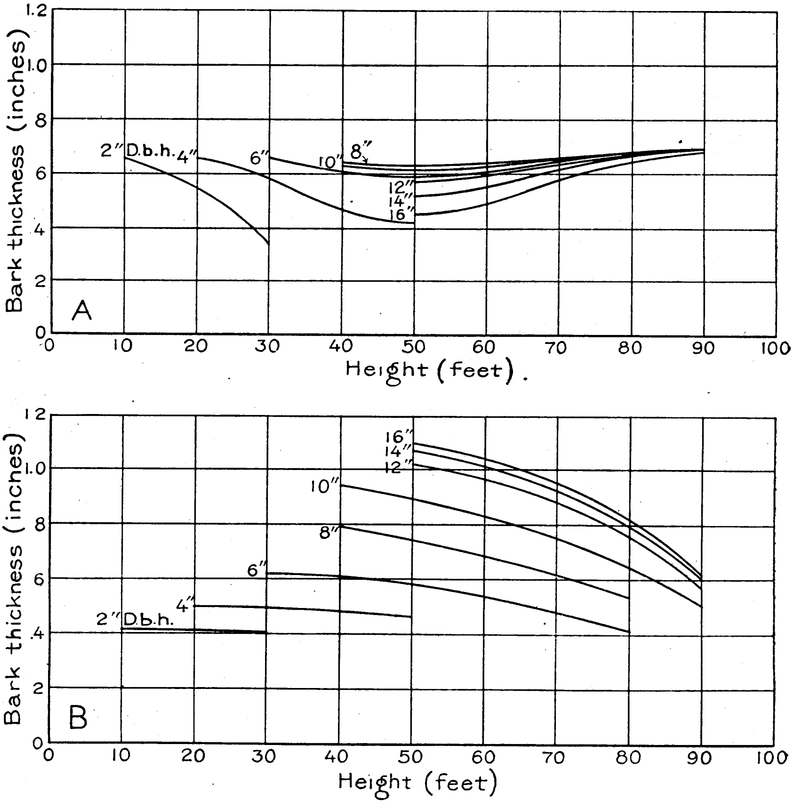

FIGURE 32.—A, Curves of bark thickness resulting from use of assumed alinement chart; B, curves of bark thickness resulting from use of correlation alinement chart

The radical nature of these changes suggests the possibility of starting with an assumed chart, thus eliminating the linear correlation step. Varying success will attend such procedure, and the validity of the final results can seldom be checked except by comparing with those derived from a chart based on a multiple regression equation. There is, then, little to be gained by such procedure. The possibility of serious error from the use of haphazardly assumed charts can be illustrated by the following example, in which bark thickness at stump is correlated with breast-high diameter and total height. A multiple regression equation was computed to be—

$$\text{Bark thickness} = 0.0598 \times \text{d. b. h.} - 0.0054 \times \text{height} + 0.481$$

An assumed chart was prepared with the height scale purposely made positive in sign to test the method. A comparison of the final curves shown in Figure 32, A and B, will show that those originating in a correlation are far more reasonable than those obtained with the assumed chart. The high intercorrelation between breast-high diameter and total height, together with the relatively small variation with height, combine to make difficult the definition of the correct relationship.

INITIAL CORRECTIONS

In each of the illustrative examples given, the first correction for curvature was made for the independent variables. It has been stated that a quicker development and stabilization of curvature sometimes results when the dependent variable is corrected for curvature first. This was true in a study defining the relation between board-foot and cubic-foot ratio of trees and their breast-high diameter and total height. In one such study the number of estimates necessary was reduced from five to three by correcting the dependent variable first. Quite naturally, also, each of the curves was better defined than those secured when the independent variables were corrected first. Unfortunately, it is not possible, without previous experience with problems of the same kind, to tell which type of correction should be applied first.

USE OF KNOWLEDGE OF CURVE FORM

In connection with another study of the relation between the board foot—cubic-foot ratios and breast-high diameter and total height, an interesting example occurred of the use of knowledge of curve form. It can be deduced from a comparison of the board-foot contents of logs with their cubic-foot volumes that the board foot-cubic foot ratio will increase rapidly with an increase in the small diameters, and less rapidly with an equal increase in the large diameters, becoming nearly constant for very large logs but in no case showing a decrease. This should also hold true for trees, except that the ratio will be slightly lower because of the small top log always present and because of the top itself, which has no board-foot volume.

First-estimate residuals for diameter breast high as computed and plotted about the original graduating curve are shown in Figure 33. The curve indicated has a sag in the center, and therein is unlike what the true curve is known to be. Accordingly the fitted curve (though correctly balanced) ignores this sag and assumes a continuously rising curve of regularly decreasing slope.

By ignoring this sag the final curves were obtained with only five estimates, as against the seven required when the sag was incorporated.

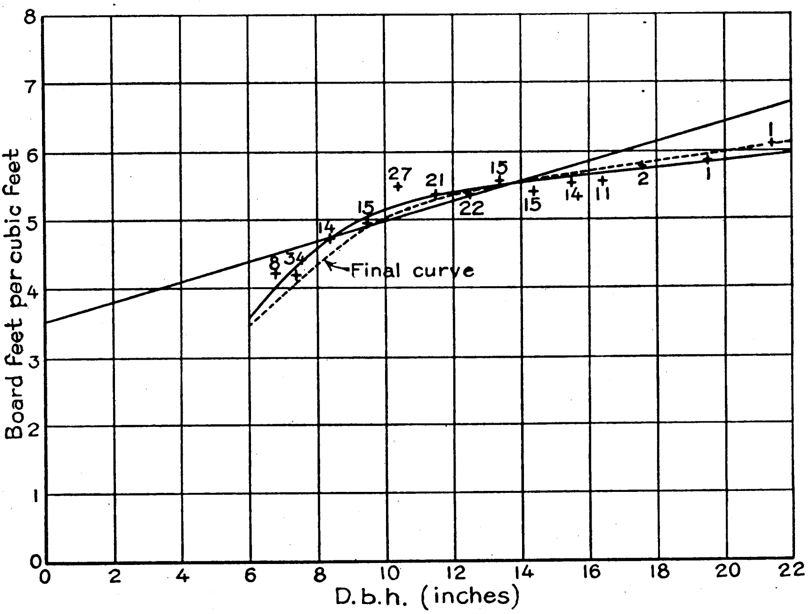

FIGURE 33.—Because of what is known of curve form the sag in the regression line indicated by the residuals over d. b. h. has been ignored, and a continuously rising, balanced curve was drawn through the points. The correctness of this treatment is confirmed by the final curve

RESIDUALS IN TERMS OF AN AUXILIARY REGULAR SCALE

In the first problem presented in this bulletin, residuals were measured in two ways—(1), in terms of the dependent variable itself (differences between measured and estimated values), and (2) for the final correction, in terms of an auxiliary regular scale. This second type of measurement was necessary, as was explained, because of the variable spacing of the final graduations for the dependent variable, which would not permit adding the correcting regression equation to the final chart. Similar situations occur wherever the dependent variable scale is modified after an estimate, and the measurement of all residuals by a regular scale might seem advisable. In some problems such a procedure is advisable and assures better results. In other cases the reverse is true, as in the dextrose problem. This is shown in Table 11, which lists the standard errors resulting from the use of each of the two types of measurement of residuals. Obviously, the better results were got when the residuals were measured in terms of the dependent variable. This table also lists the standard errors resulting from both types of treatment of the first problem. Here, the measurement of the residuals in terms of an auxiliary regular scale gave decidedly better results.

TABLE 11.—*Comparison of standard errors of successive estimates in problems in which residuals were measured (1) in terms of the dependent variable, and (2) in terms of an auxiliary, regular scale*

A. DEXTROSE PROBLEM

Residuals in terms of—	Standard errors of successive estimates							
	First	Second	Third	Fourth	Fifth	Sixth	Seventh	Eighth
Dependent variable	13.94	12.17	6.22	5.67	5.56	5.42	5.42	----
Regular scale [1]	13.94	12.23	5.81	8.07	7.23	6.89	7.15	·7.24

B. PROBLEM (page 22): $\left(\frac{Z}{4}\right)^2=\sqrt{X}+10\log Y$

Dependent variable	1.14	.50	.48	.36	.33	.28	.27	.24
Regular scale	1.14	.50	.48	.36	.29	.22	.21	----

[1] Note the increase in standard error in the fourth, seventh, and eighth estimates.

A third course is to measure, not individual deviations, but class-average deviations in terms of the auxiliary regular scale. This may be particularly desirable when using correction distances, since the correction distance is a constant, times the length of scale covered by one unit of the dependent variable. When this unit length is changed by modifying the scale, the correction distance changes also. It is more logical, therefore, when using correction distances, to measure the residuals in terms of the scale unit on which the correction distances are based, although, as indicated by the footnote on page 53, better results may often be obtained by applying correction distances to the residuals in terms of the dependent variable.

Any convenient scale may be used for measuring these residuals, but since the original scale of the chart is regular and already at hand it may be more desirable to use it in preference to another. The first-estimate residuals in terms of this regular scale are then identical with the residuals in terms of the dependent variable.

No method has been discovered of foretelling which treatment is superior. The only possible procedure is to adopt one at random and watch the results closely. If the improvement is unduly small, especially when the curvature is great, one of the others should be substituted.

FIELD OF APPLICATION

The field of application of the method presented in these pages is wide. It includes all branches of forest research in which quantitative measurements are made, wherever more than two variables are involved, and wherever curvilinear relations are suspected. It would be impossible to list all or even a large proportion of the problems which may be solved by this technic, and it must suffice to point out a few scattered types.

The various examples already illustrated show certain applications; possibilities exist for applying it to physiological studies, in associating cell-sap density or osmotic pressure with site quality, soil and air temperatures, soil moisture, and position of samples in trees in absolute or percentage units of height; resin flow may be related to size and age of tree and to climatic factors; viability of seed, as affected by size, weight, age, and storage temperatures, may be studied; germinative capacity of seed may be related to size and age of tree and site quality, or to length of storage, depth of planting, rainfall, and temperature.

In nursery practice, size of plants may be related to amount of fertilizer or other chemicals, amount of water, and growing space.

In fire studies the method may be useful in associating rate of spread with atmospheric and fuel conditions, or in predicting inflammability (fuel moisture) from atmospheric conditions.

Skidding time, or cost, in logging studies may be correlated with log diameter, length, and skidding distance, or sawing time in the mill may be correlated with size, amount of defect, etc.

In pulp and paper investigations, yields or breaking strength may be related to composition of liquor, temperature, pressure, duration of cook, percentages of species or types of pulp, etc.

Correlation methods have been used so much in economic studies that the place of this method in forest economics is obvious. Many economics problems, however, require the correlation of time series, which, as noted before, require extreme care.

When to use the method can be decided only after consideration of the kind and characteristics of the data available. In 3-variable problems, where the curves are similar in form, and where there is no intercorrelation between the independent variables, as in yield tables, the use of anamorphosis or other analogous methods will give results more quickly.

In connection with controlled experiments, the value of the method is not always sufficient to warrant its use. If the results are erratic, however, it may pay to use it as a smoothing medium, as in the dextrose problem, and for ease of estimating, as well as for interpreting the results.

CRITERIA OF APPLICABILITY

The criteria by which to judge the possibilities for solution of any particular problem by this method are not rigid. The initial alienation coefficient gives some indication, but must not be given too much weight. For instance, in a taper-curve problem, obtaining diameters at any point from percentage height of the point, diameter breast high, and total height, the alienation coefficient was only 0.354. This was low, but the regression equation represented the best-fitting cones, and obviously was not a good fit of the data, although, as a starting point for further investigation, it is quite satisfactory.

On the other hand, a high alienation coefficient between the variables may not prevent a satisfactory solution. In the construction of a cubic-foot volume table from correlated form factors the alienation coefficient was 0.929 and the final alienation index was only reduced to 0.909. The resulting volume table was satisfactory because the range of form factor values was very small ($SD=0.0432$). Since the volume of a tree is determined chiefly by its height and diameter (volume = height × basal area × form factor) a small variation in form factor produces only a small change in volume. In other words, the form factors are used here merely to compute volumes and have no other significance. It is proper, therefore, to measure accomplishment by the final alienation index in volume rather than in form factor. If this is done a value which is very much lower (approximately 0.099, instead of 0.909 for form factor) is obtained.

Experience and good judgment will help a great deal to decide what procedure should be followed, and good judgment must be exercised at all times to avoid dogmatic acceptance of unreasonable results. The nature of such results may suggest the possibility of using cumulations, ratios, logarithms, etc., which may very possibly clear the situation. Converting the original data and working entirely with logarithms may give the proper answer in those cases where deviations are more nearly constant percentages, rather than constant absolute values.

SUMMARY

Purely graphic methods are inadequate for the solution of many forestry problems. They are too dependent on the judgment of the investigator, have led to serious errors, and as ordinarily applied yield results which are unchecked as to accuracy and unappraised as to accomplishment. They are inapplicable to cases involving more than three variables, and even where three are involved they demand enormous numbers of data in order to produce fairly satisfactory results.

The concepts of the modern science of statistics may be applied to graphs and curves. By this means a large gain in accuracy over the familiar graphic processes may be obtained. The common statistical processes, however, are too rigid in their assumptions to be useful in many forestry problems.

The curvilinear-correlation method combines graphic and statistical technic. The former contributes flexibility, the latter accuracy. The basic assumptions are so generalized that a very wide range of problems may be solved by it, yet the results are rigorously checked and appraised. It permits a solution of many problems previously considered insoluble on account of complexity or number of data required.

The computations seem laborious, but the extra work involved in the office is often fully compensated by the reduction in the number of field data. Where this compensation is only partial the greater total labor of the new method is fully repaid by the better results obtained and by an accurate knowledge of what has been accomplished.

The technic has numerous variations. A careful choice should be made among them, dependent on the peculiarities of the problem being studied. It follows that this technic does not lend itself to mechanical handling. Intelligent alertness is an essential to success.

The final result is in the form of an alinement chart. While a table or tables may and doubtless will be read from this, the chart is an exceedingly compact method of presenting a complex relation. It is particularly convenient where many interpolations would be necessary if the tabular form were substituted.

APPENDIX

SHORT-CUT METHODS

STATISTICAL MEASURES

Many of the computations involved in the methods which have been discussed are formidably laborious, particularly where large numbers of data are used. Fortunately, numerous short-cut methods have been devised. These have not previously been described because to do so would have merely confused the reader who had not yet gained an adequate comprehension of the principles involved. The formulae used heretofore are simpler to understand, while the short-cut formulae are easier to apply.

It is not necessary to list here all the abridged methods which have been devised, and only those which are most useful to forestry workers will be given. Three of these are based on the following identities—

$$\text{Sum } d^2_X = \text{Sum } X^2 - NM^2_X$$

$$= \frac{N(\text{Sum } X^2) - \text{Sum}^2\ X}{N} \text{ -------------- (XIV)}$$

and—

$$\text{Sum } (d_X d_Y) = \text{Sum } XY - NM_X M_Y$$

$$= \frac{N(\text{Sum } XY) - (\text{Sum } X)(\text{Sum } Y)}{\text{N}} \text{ -----------(XV)}$$

where, as before, X and Y are two variables, M_X is the mean value of X, M_Y is the mean value of Y, while d_X and d_Y signify the deviations of the values of X and Y, respectively, from their means. The formulae which have previously been used for standard deviation, alienation coefficient, and the regression equation may, therefore, be modified as follows—

Standard deviation (p. 7)—

$$SD_X = \sqrt{\frac{\text{Sum } d^2_X}{N}} \text{ -------------------------(II)}$$

Substituting from equation XIV—

$$SD_X = \frac{\sqrt{N(\text{Sum } X^2) - \text{Sum}^2\ X}}{N} \text{ ----------------- (XVI)}$$

Regression equation (p. 10)—

$$Y = M_Y + \frac{\text{Sum } d_X d_Y}{\text{Sum } d^2_X}(X - M_X) \text{ ------------------(V)}$$

Substituting from formulae XIV and XV—

$$Y = M_Y + \frac{N(\text{Sum } XY) - (\text{Sum } X)(\text{Sum } Y)}{N(\text{Sum } X^2) - \text{Sum}^2\ X}(X - M_X) \text{ ------(XVII)}$$

Alienation coefficient (p. 12)—

$$AC_{XY} = \sqrt{1 - \frac{\text{Sum}^2\ (d_X d_Y)}{(\text{Sum } d^2_X)(\text{Sum } d^2_Y)}} \text{ ---------------(VI)}$$

Substituting from formulae XIV and XV—

$$AC_{XY} = \sqrt{1 - \frac{[N(\text{Sum } XY) - (\text{Sum } X)(\text{Sum } Y)]^2}{[N(\text{Sum } X^2) - \text{Sum}^2 X][N(\text{Sum } Y^2) - \text{Sum}^2 Y]}} \text{ ---(XVIII)}$$

Correlation coefficient (p. 8)—

$$CC_{XY} = \sqrt{1 - (AC_{XY})^2} \text{ --(IV)}$$

$$= \frac{N(\text{Sum } XY) - (\text{Sum } X)(\text{Sum } Y)}{\sqrt{[N(\text{Sum } X^2) - \text{Sum}^2 X][N(\text{Sum } Y^2) - \text{Sum}^2 Y]}} \text{ --------(XIX)}$$

It will be observed that the revised equations XVI to XIX contain but a few basic quantities, which are combined in different ways. This means that a single calculation can be performed which will yield the basic values necessary for the computation of standard deviation, alienation coefficient, and the regression equation. These basic values are the sums of each variable, their means, the sums of the squared variables, and the sums of their products when multiplied together in pairs.

As an example, the new formula will be used in connection with the material presented in Table 3. Table 12 illustrates the short-cut method.

TABLE 12.—*Short-cut method of calculating standard deviations, regression equation, and alienation coefficient*

Age—(X)	Measured d.b.h.—(Y)	XY	X^2	Y^2
(1)	(2)	(3)	(4)	(5)
15	6.0	90.0	225	36.00
19	9.7	184.3	361	94.09
22	10.5	231.0	484	110.25
25	14.0	350.0	625	196.00
29	12.7	368.3	841	161.29
34	16.5	561.0	1,156	272.25
43	18 0	774.0	1,849	324.00
45	19.5	877.5	2,025	380.25
52	21.0	1,092.0	2,704	441.00
55	19.0	1,045.0	3,025	361.00
58	22.5	1,305.0	3,364	506.25
65	24.0	1,560.0	4,225	576.00
66	22.1	1,458.6	4,356	488.41
75	22.0	1,650.0	5,625	484.00
75	24.0	1,800.0	5,625	576.00
81	25.0	2,025.0	6,561	625.00
88	25.0	2,200.0	7,744	625.00
89	23.5	2,091.5	7,921	552.25
91	25.1	2,284.1	8,281	630.01
100	24.5	2,450.0	10,000	600.25
1,127	384.6	24,397.3	76,997	8,039.30
56.35	19.23	1,219.86	---------	-----------

$$\left.\begin{aligned} SD_X&=\frac{\sqrt{20(76997)-(1127)^2}}{20}=25.972 \\ SD_Y&=\frac{\sqrt{20(8039.30)-(384.6)^2}}{20}=5.672 \end{aligned}\right\}\text{-----------------------} \quad \text{(XVI)}$$

$$Y=19.23+\frac{(20)(24397.3)-(1127)(384.6)}{(20)(76997)-(1127)^2}(X-56.35)\text{--------}\text{(XVII)}$$

$$=19.23+0.202\ (X-56.35)$$

$$=0.202X+7.85$$

$$AC_{XY}=\sqrt{1-\frac{[(20)(24397.3)-(1127)(384.6)]^2}{[(20)(76997)-(1127)^2][(20)(8039.30)-(384.6)^2]}}\text{--}\text{(XVIII)}$$

$$=\sqrt{1-0.8555}$$

$$=0.380$$

Where these short-cut methods [27] are used, the standard errors, coefficients of correlation, etc., may readily be derived from the standard deviations and alienation coefficients.

The time saved by this procedure is obviously considerable, since the individual deviations need not be computed. The slightly greater complexity of the formulas themselves is negligible where any considerable number of data are involved. The seven columns of Table 3 have been reduced to the five columns of Table 12. To offset this advantage, in part, the actual figures handled have become larger. Furthermore, the number of significant figures which must be retained has increased. It will be noted that each of the new formulæ involves the subtraction of one large number from another in one or more instances. These large numbers may be nearly the same in value, and where this is the case, one or more of the left-hand digits disappears. In addition, the process of extracting the square root cuts the number of significant figures about in two. As a result, it is often necessary to retain six or seven significant figures in the early parts of the computation in order to insure accuracy to two or three figures in the values sought. The slide rule is therefore not suitable for work of this nature, and a good calculating machine is almost essential to efficiency. Where very large numbers of data are involved, a great saving in time can be made by the use of electrical tabulating machinery.

The use of machines effects a considerable saving in the computation of regression equations. If a calculating machine with the carry-over feature in the result

[27] Another short-cut method of computing standard deviations is described in Croxton (*6*).

dial is available, as in certain models of the best-known makes, the sums of X, X^2, and XY, or similar combinations can be obtained at one operation.

To do this put Y in the left-hand side of the keyboard, X in the right-hand side, and multiply by X, thus entering $Y \times X$ and $X \times X$ or X^2 in the carriage dial, and X in the result dial. Do not clear these dials. Enter each item in the same manner and when all items have been thus entered the carriage dial will contain Sum $(Y \times X)$ and Sum (X^2), and the result dial will show Sum (X).

For a 3-variable problem involving X, Y, and Z, three such sets of sums are required. These may be taken for—

(1) Sum (X), Sum (X^2), and (first run), Sum (XY), or (check run) Sum (XZ);
(2) Sum (Y), Sum (Y^2), and (first run), Sum (YZ), or (check run) Sum (YX);
(3) Sum (Z), Sum (Z^2), and (first run), Sum (ZX), or (check run) Sum (ZY).

PUNCHED-CARD TABULATING EQUIPMENT

Those who have punched-card tabulating equipment (*34, p. 94*) available will be able to materially reduce the amount of labor involved in computing the regression equation when a large amount of data are involved. About 200 to 250 items, to be sorted at least three times, is the smallest amount which will be handled by this equipment with any saving in time. Large groups of data, especially if more than three variables are involved, are handled at a very great saving of time and effort.

A special technic for using this equipment in correlation problems has been devised.[28] This technic calls for the used of coded values, thereby permitting the use of automatic checks throughout the progress of the computations.

CODING

Coding is frequently used in mechanical and other methods to reduce the size of large numbers, as well as to permit the use of the checking system mentioned above.

By subtraction, division, or both, and rounding off, the values are reduced to a small series of whole numbers, preferably ranging from 0 to about 15. A range as small as from 0 to 10 or as large as from 0 to 30 is satisfactory. Coding is essentially a grouping of the values into numbered classes in such a way that the grouping may be expressed algebraically. For example, in the bark-thickness problem on page 54, diameter ranged from 4 to 15 inches, covering 12 inch-classes. These numbers may be reduced in size by subtracting four from each class, the coded classes then ranging from 0 to 11. The coding may be expressed—

$$\text{Coded d. b. h.} = \text{d. b. h. (nearest inch)} - 4.$$

Total height ranged from 25 to 90. When each height is reduced by 25, the range becomes 0 to 65. This can further be divided into sixteen 4-foot classes. The coding can be expressed—

$$\text{Coded total height} = \frac{\text{Total height} - 25}{4}$$

Easily handled subtrahends and divisions, for easy mental calculation of coded values, should be used. The accuracy of the work is in no wise affected when data are coded by subtraction. Accuracy, however, is affected to some extent by coding by division, the accuracy decreasing as the divisor increases. The errors involved are similar to those resulting from measuring, for example, diameters to the nearest inch, or nearest even inch. The inaccuracies involved are not great, but should be given consideration in preparing the data. In no case should there be less than 10 coded classes for any variable.

AVERAGE ERROR

When the data are numerous enough to justify the use of special equipment it is also justifiable, in general, to use the average error rather than the standard error to follow the improvement with each approximation. The following short cut may be used in computing this value.

Total separately the measured and estimated values of all items for which the measured value exceeds or equals [29] the estimated value. The difference between these totals is the sum of the positive deviations.

[28] SMITH, B. B. THE USE OF PUNCHED-CARD TABULATING EQUIPMENT IN MULTIPLE CORRELATION PROBLEMS. U. S. Dept. of Agr. Bur. Agr. Econ. 1923. [Mimeographed.]

[29] Those which are equal could be omitted entirely or included in the next step, without affecting the average error in absolute units, but are necessary if the average error or aggregate deviation is to be computed as a percentage of the mean value of the variable.

Repeat for items in which the measured values are less than the estimated. The sum of the negative deviations is thus obtained.

The sum of the positive and negative deviations, disregarding signs, divided by the total number of items is the average error (*AE*). The two totals of measured values obtained above may be summed to obtain an aggregate of the measured values, and similarly the aggregate of estimated values may be obtained. Aggregate estimated minus aggregate measured gives the aggregate deviation (observe sign) in the units used. This value, divided by the aggregate estimate and multiplied by 100 gives the aggregate deviation per cent.

AVERAGE DEVIATION FROM THE MEAN

A method somewhat analogous to that used for determining the average error may be used to compute the average deviation (*AD*) as follows:

Determine the mean. Sum the items greater than the mean. Multiply the number of such items by the mean and subtract the product from the sum determined above. Double the difference and divide by the total number to obtain the average deviation.

SYMBOLS

There is no universally accepted system of symbols for statistical measures. Systems most widely used have employed Greek letters, but these are inconvenient where the study is to be written up on the typewriter. Throughout this bulletin initial letters have served as symbols. Their advantages are that they are easily remembered and that they can be employed in typewritten manuscripts without inconvenience.

The symbols employed in this bulletin are listed below with their common equivalents. Alternative designations are given in parenthesis.

Symbol	Designation	Equivalent
AC	Alienation coefficient	k.
AC_{XY} AC_{YX}	Alienation coefficient between X and Y	k_{XY}. k_{YX}.
AD	Average deviation	
AE	Average error	
AI	Alienation index	
B_{XY}	Partial regression coefficient of X on Y	β_{XY}.
B_{YX}	Partial regression coefficient of Y on X	β_{YX}.
CC	Correlation coefficient	r.
CC_{XY} CC_{YX}	Correlation coefficient between X and Y	r_{XY}. r_{YX}.
CI	Correlation index	P.
CI_{XY} CI_{YX}	Correlation index between X and Y	P_{XY}. P_{YX}.
d	Deviation from mean	
d_X	Deviation of X from mean of X	
e	Residual (error, or deviation of individual item from curve)	
e_1, e_2	Residual of first, second estimate	
e_X, $e_{f(x)}$	Residual expressed in terms of X, in terms of function of X	
M	Arithmetic mean	
M_X	Arithmetic mean of X	
N	Number of items	n.
SD	Standard deviation	σ or s.
SD_X	Standard deviation of X	σ_X or s_X.
SD_M	Standard deviation of mean (standard error)	S.
SE	Standard error (standard deviation about curve)	
Sum( )	Sum of all values of symbol which follows	$\Sigma(\ )$.
$\text{Sum}^2(\ )$	Squared sum of all values of symbol which follows	$\Sigma^2(\ )$.
$\text{Sum}(\ )^2$	Sum of squares of all values of symbol which follows	$\Sigma(\ )^2$.

FORMULAE

Standard error (p. 6)—

$$SE=\sqrt{\frac{\text{Sum } e^2}{N}} \quad \text{------------------------------} \quad \text{(I)}$$

Standard deviation (p. 7)—

$$SD=\sqrt{\frac{\text{Sum } d^2}{N}} \quad \text{------------------------------} \quad \text{(II)}$$

Alienation index (p. 8)—

$$AI=\frac{SE}{SD} \quad \text{------------------------------} \quad \text{(III)}$$

Correlation index (p. 8)—

$$CI=\sqrt{1-(AI)^2} \tag{IV}$$

Regression equation (p. 10)—

$$Y=M_Y+\frac{\text{Sum } d_X d_Y}{\text{Sum } d^2{}_X}(X-M_X) \tag{V}$$

Alienation coefficient (p. 12)—

$$AC_{XY}=\sqrt{1-\frac{\text{Sum}^2\ d_X d_Y}{(\text{Sum } d^2{}_X)\ (\text{Sum } d^2{}_Y)}} \tag{VI}$$

Multiple regression equation (p. 16)—

$$W=M_W+B_{WX}\frac{SD_W}{SD_X}(X-M_X)+B_{WY}\frac{SD_W}{SD_Y}(Y-M_Y) \tag{VII}$$

Normal equations—three variables (p. 17)—

$$\left.\begin{aligned} B_{WX}+CC_{XY}B_{WY}&=CC_{WX}\\ CC_{YX}B_{WX}+B_{WY}&=CC_{WY}\end{aligned}\right\} \tag{VIII–A}$$

Normal equations—four variables (p. 17)—

$$\left.\begin{aligned} B_{WX}+CC_{XY}B_{WY}+CC_{XZ}B_{WZ}&=CC_{WX}\\ CC_{YX}B_{WX}+B_{WY}+CC_{YZ}B_{WZ}&=CC_{WY}\\ CC_{ZX}B_{WX}+CC_{ZY}B_{WY}+B_{WZ}&=CC_{WZ}\end{aligned}\right\} \tag{VIII–B}$$

Alienation coefficient (p. 18)—

$$AC_{W(XYZ\,-\,-)}=\sqrt{1-(B_{WX}CC_{WX}+B_{WY}CC_{WY}+B_{WZ}CC_{WZ}+\,-\,-\,-\,-)} \tag{IX}$$

Alinement chart formulae (pp. 41–42)—

$$U=Lf(\quad) \tag{X}$$

$$L_Z=\frac{L_X L_Y}{L_X+L_Y} \tag{XI}$$

$$\frac{XZ}{ZY}=\frac{L_X}{L_Y} \tag{XII}$$

Correction distances (p. 50)—

$$\text{cor. } X=\frac{XY}{ZY}L_Z \tag{XIII–A}$$

$$\text{cor. Y}=\frac{XY}{XZ}L_Z \tag{XIII–B}$$

Shortcut formulae (pp. 76)—

$$\text{Sum d}^2{}_X=\frac{N(\text{Sum } X^2)-\text{Sum}^2X}{N} \tag{XIV}$$

$$\text{Sum } (d_X d_Y)=\frac{N(\text{Sum } XY)-(\text{Sum } X)(\text{Sum } Y)}{N} \tag{XV}$$

$$SD_X=\frac{\sqrt{N(\text{Sum } X^2)-\text{Sum}^2X}}{N} \tag{XVI}$$

Regression equation (p. 76)—

$$Y=M_Y+\frac{N(\text{Sum } XY)-(\text{Sum } X)\ (\text{Sum } Y)}{N(\text{Sum } X^2)-\text{Sum}^2\ X}(X-M_X) \tag{XVII}$$

$$AC_{XY}=\sqrt{1-\frac{[N(\text{Sum } XY)-(\text{Sum } X)\ (\text{Sum } Y)]^2}{[N(\text{Sum } X^2)-\text{Sum}^2\ X]\ [N(\text{Sum } Y^2)-\text{Sum }{}^2Y]}} \tag{XVIII}$$

$$CC_{XY}=\frac{N(\text{Sum } XY)-(\text{Sum } X)\ (\text{Sum } Y)}{\sqrt{[N(\text{Sum X}^2)-\text{Sum}^2X]\ [N(\text{Sum } Y^2)-\text{Sum}^2Y]}} \tag{XIX}$$

TABLE 13.—*A list of representative regression equations and their associated statistical measures*

LABORATORY TESTS—1 GRAM DEXTROSE IN 25 CUBIC CENTIMETERS SULPHURIC ACID FOR 30 MINUTES

Dependent variable	Independent variables and regression coefficients	Constant	Number of items	Mean of dependent variable	SD of dependent variable	Estimate number	SE of dependent variable	AC	AI
Residual dextrose (per cent)	−12.7556 sulphuric acid (per cent) −1.7431 temperature (° C.)	+378.4227	32	61.58	33.13	7	5.42	0.421	0.164

LABORATORY TESTS ON LONGLEAF PINE

Dependent variable	Independent variables and regression coefficients	Constant	Number of items	Mean of dependent variable	SD of dependent variable	Estimate number	SE of dependent variable	AC	AI
Yield of crude pulp (per cent)	−0.4742 Na_2O (pounds per 100 pounds of chips) −0.2306 steam pressure (pounds per square inch gauge) −2.1274 total time of cooking (hours)	+95.2686	73	50.51	12.44	1	6.04	0.486	--------

CONIFEROUS SEEDLINGS IN NURSERY

Dependent variable	Independent variables and regression coefficients	Constant	Number of items	Mean of dependent variable	SD of dependent variable	Estimate number	SE of dependent variable	AC	AI
Seedlings damped-off per hour (per cent of initial number).	−0.0469 age (days) −0.0069 soil temperature (° E.) +0.0288 evaporation (cubic centimeter per hour) −0.0039 soil moisture, top 0.6 centimeter (per cent)	+1.4713	[1]	0.11	0.119	6	0.058	0.785	0.485

SECOND-GROWTH SHORTLEAF PINE STEM MEASUREMENTS

Dependent variable	Independent variables and regression coefficients	Constant	Number of items	Mean of dependent variable	SD of dependent variable	Estimate number	SE of dependent variable	AC	AI
Single bark thickness (inches)	+0.06263 d. b. h. (inch) of tree −0.00752 height (foot) of tree +0.00519 age (year) of tree −0.02068 d. i. b. (inch) of section −0.000941 height (foot) of section −0.009873 height (per cent of total) of section −0.00447 age (years) of section +0.00078 site index (height in feet at 50 years)	+0.75076	[2] 515	0.382	0.300	1	0.172	0.575	--------
Single bark thickness (inches)	+0.07034 d. b. h. (inch) of tre −0.00600 height (foot) of tree −0.03324 d. i. b. of section −0.00978 height (per cent of total) section	+.78716	[2] 515	.382	.300	1	.173	.577	--------
						2	.129	----------	0.430
						3	.124	----------	.413
						4	.123	----------	.410
						5	.121	----------	.403

[1] Records for 32 days.

[2] 124 trees.

TABLE 13.—*A list of representative regression equations and their associated statistical measures*—Continued

SECOND-GROWTH SHORTLEAF PINE STEM MEASUREMENTS—Continued

Dependent variable	Independent variables and regression coefficients	Constant	Number of items	Mean of dependent variable	SD of dependent variable	Estimate number	SE of dependent variable	AC	AI
Single bark thickness residuals, fifth estimate (inches).	+0.000400 d. b. h. of tree −0.0000862 height of tree −0.00162 d. i. b. of section −0.000132 height (per cent of total) section	+.0174	[2] 515	.069	.123	1	.1175	.955	--------
Single bark thickness (inch)	+0.04979 d. b. h. of tree −0.00666 height of tree −0.00734 height (per cent of total) of section	+.70228	[2] 515	.382	.300	1	.177	.590	--------
Single bark thickness (inch)	+0.00364 d. b. h. of tree +0.00771 height of tree +0.03717 d. i. b. of section	−.31594	[2] 515	.382	.300	1	.268	.893	--------

SECOND-GROWTH LOBLOLLY PINE STEM MEASUREMENTS

Dependent variable	Independent variables and regression coefficients	Constant	Number of items	Mean of dependent variable	SD of dependent variable	Estimate number	SE of dependent variable	AC	AI
Stump bark thickness (inches)	+0.0615 stump d. i. b. −0.0033 total height	+0.4585	[3] 355	0.795	0.266	1	0.218	0.820	--------
Stump d. i. b. (inches)	+0.948 d. b. h. +0.0147 total height	−.32	[3] 355	8.87	3.36	1	1.06	.315	--------
Bark thickness at half height (inches).	+0.0384 d. i. b. at one-half height −0.0024 total height	+.1686	[3] 355	.23	.091	1	.076	.835	--------
D. i. b. at one-half height (inch)	+0.6060 d. b. h. +0.0081 total height	−.2657	[3] 355	5.53	2.13	1	.482	.226	--------
D. i. b. (inches)	+0.5873 d. b. h. of tree +0.0138 height of tree +0.0121 age of tree −0.0734 height (per cent of total) of section	+2.2141	[2] 515	5.56	3.02	1	1.07	.355	--------
D. i. b. (inches)	+0.5876 d. b. h. of tree +0.0239 height of tree −0.0733 height (per cent of total) of section	+2.5793	[2] 515	5.56	3.02	1	1.07	.355	--------

SECOND-GROWTH SLASH PINE

D. b. h.	−7.5900 form factor (cylinder) +.1694 height (total)	+1.1892	³ 269	9.85	3.62	1	1.835	0.507	--------
Total height	+4.3205 d. b. h. +69.650 form factor (cylinder)	−1.2091	³ 269	68.77	16.29	1 3 5	3.36 2.47 2.29	.206 ---------- ----------	-------- 0.152 .141
Form factor	−.00404 d. b. h. +.00145 height (total)	+.33350	³ 269	.394	.0432	1 2 3 4 5	.032 .031 (AD).030 .030 .030	.741	--------
From factor (residuals)	+.001185 d. b. h. −.000154 height (total)	−.001138	³ 269	.119	.0382	1	.03819	.9997	--------
Volume, entire stem less bark (cubic feet).	+3.9805 d. b. h. +.0758 height (total)	−25.4615	³ 269	18.98	16.29	1	4.72	.290	--------
Do.	+31.2800 basal area (square feet) +.1621 height (total)	−10.9545	³ 269	18.98	16.29	1	3.34	.205	--------
Board feet (Int. ⅛-in.) per cubic foot.	+.14341 d. b. h. +.014563 height (total)	+0.02854	³ 269	5.15	1.169	⁴ 1	.523 .379	.447 ----------	-------- .324

SECOND-GROWTH SHORTLEAF PINE

Board feet (Int. ⅛-in.) per cubic foot.	+0.1511 d. b. h. + .0748 height (total)	−1.8038	³ 227	4.73	1.94	⁴ 1	1.38 .706	0.711 ----------	-------- 0.364

SECOND-GROWTH LOBLOLLY PINE

Board feet (Int. ⅛-in.) per cubic foot.	+0.1973 d. b. h. + .0612 height (total)	−2.0595	³ 294	4.31	2.16	1	1.24	0.574	--------

SECOND-GROWTH LONGLEAF PINE

Board feet (Int. ⅛-in.) per cubic foot.	+0.2609 d. b. h. + .0382 height (total)	−0.4634	³ 377	4.51	1.23	1	0.545 (AE) .666	0.443 ----------	-------- --------
Board feet (Doyle) per cubic foot.	+0.3158 d. b. h. + .0203 height (total)	−3.2313	³ 377	3.75	1.61	1	.445 (AE) .322	.276 ----------	-------- --------

[2] 124 Trees.

[3] Trees.

[4] Final.

TABLE 13.—*A list of representative regression equations and their associated statistical measures*—Continued

SECOND-GROWTH WESTERN WHITE PINE, YIELD AND VOLUME MEASUREMENTS

Dependent variable	Independent variables and regression coefficients	Constant	Number of items	Mean of dependent variable	SD of dependent variable	Estimate number	SE of dependent variable	AC	AI
Form factor (cylinder)	−0. 00642 d. b. h. of trees + . 00001 total height of trees + . 00005 site index of stand + . 00013 age of stand + . 00007 density of stand (per cent of normal BA) of stand + . 00003 per cent of white pine (BA) of stand	+0. 48045	[3] 649	0. 437	0. 049	1	0. 0397	0. 810	--------
Do	− . 00668 d. b. h. of tree + . 00009 height of tree + . 00008 site index of stand + . 00009 per cent white pine (BA) of stand + . 00026 per cent tolerant species (BA) of stand	+. 48911	[3] 649	. 437	. 049	1	. 0398	. 812	--------
Do	− . 00669 d. b. h. of tree + . 00009 height of tree + . 00010 per cent white pine (BA) of stand + . 00026 per cent tolerant species (BA) of stand	+. 49309	[3] 649	. 437	. 049	1	. 0398	. 812	--------

MATURE WESTERN YELLOW PINE

Dependent variable	Independent variables and regression coefficients	Constant	Number of items	Mean of dependent variable	SD of dependent variable	Estimate number	SE of dependent variable	AC	AI
Frustum form factor (average by locality)	+0. 0112 average d. b. h. − . 0001 average merchantable length − . 0010 site index (Merchantable height at maturity)	+0. 8392	[5] 493	0. 990	0. 188	1	0. 148	0. 787	--------

[3] Trees

[5] Locality averages (11,276 trees).

PARTIAL LIST OF STATISTICAL LITERATURE [28]

BOOKS AND ARTICLES

(1) Bowley, A. L.
1920. ELEMENTS OF STATISTICS. Ed. 4, 459 p. illus. London.

(2) ———
1928. THE STANDARD DEVIATION OF THE CORRELATION COEFFICIENT. Jour. Amer. Statis. Assoc. 23 (n. s. 161): 31–34.

(3) *Bruce, D.
1919. ALINEMENT CHARTS IN FOREST MENSURATION. Jour. Forestry 17: 773–801, illus.

(4) *———
1925. SOME POSSIBLE ERRORS IN THE USE OF CURVES. Jour. Agr. Research 31: 923–928, illus.

(5) Chaddock, R. E.
[1925]. PRINCIPLES AND METHODS OF STATISTICS. 471 p., illus. Boston, New York, [etc.].

(6) Croxton, F. E.
1925. AN APPARATUS TO ASSIST IN THE CALCULATION OF THE STANDARD DEVIATION OF A GROUPED FREQUENCY DISTRIBUTION. Jour. Amer. Statis. Assoc. 20 (n. s. 152): 532–536, illus.

(7) Davenport, C. B.
1904. STATISTICAL METHODS; WITH SPECIAL REFERENCE TO BIOLOGICAL VARIATIONS. Rev. ed. 2, 223 p., illus. New York.

(8) Day, E. E.
1925. STATISTICAL ANALYSIS. 459 p., illus. New York.

(9) Ezekiel, M.
1924. A METHOD OF HANDLING MULTIPLE CURVILINEAR CORRELATION FOR ANY NUMBER OF VARIABLES. Jour. Amer. Statis. Assoc. 19 (n. s. 148): [431]–453, illus.

(10) Fisher, R. A.
1925. STATISTICAL METHODS FOR RESEARCH WORKERS. 239 p., illus. Edinburgh and London.

(11) Forsyth, C. H.
1294. AN INTRODUCTION TO MATHEMATICAL ANALYSIS OF STATISTICS. 241 p., illus. New York.

(12) Gavett, G. I.
1925. A FIRST COURSE IN STATISTICAL METHOD. 358 p., illus. New York.

(13) Grasovsky, A.
1925. THE USE OF THE MEDIAN IN ESTIMATING STANDING TIMBER. Jour. Forestry 23: 71–77.

(14) Hartley, C.
1921. DAMPING-OFF IN FOREST NURSERIES. U. S. Dept. Agr. Bul. 934, 100 p., illus.

(15) Hayes, H. K.
1925. CONTROL OF SOIL HETEROGENEITY AND USE OF THE PROBABLE ERROR CONCEPT IN PLANT BREEDING STUDIES. Minn. Agr. Expt. Sta. Tech. Bul. 30, 21 p.

(16) Jenkins, T. N.
1928. APPARATUS TO FACILITATE THE CALCULATION OF THE MOMENTS OF A DISTRIBUTION. Jour. Amer. Statis. Assoc. 23 (n. s. 161): 58–60.

(17) Jones, D. C.
1921. A FIRST COURSE IN STATISTICS. 286 p., illus. London.

(18) ——— and Daniels, G. W.
1926. ELEMENT OF MATHEMATICS FOR STUDENTS OF ECONOMICS AND STATISTICS. 240 p., illus. Liverpool and London.

(19) *Karsten, K. G.
1923. CHARTS AND GRAPHS; AN INTRODUCTION TO GRAPHIC METHODS IN THE CONTROL AND ANALYSIS OF STATISTICS. 724 p., illus. New York. [Especially ch. 46 and 47.]

(20) Kelley, T. L.
1923. STATISTICAL METHOD. 390 p., illus. New York.

(21) Kincer, J. B., and Mattice, W. A.
1928. STATISTICAL CORRELATIONS OF WEATHER INFLUENCE ON CROP YIELDS. U. S. Mo. Weather Rev. 56: 53–57, illus.

[28] References marked with an asterisk are particularly valuable to foresters.

(22) King, W. I.
1918. THE ELEMENTS OF STATISTICAL METHOD. 250 p., illus. New York and London.

(23) Kressmann, F. W.
1922. THE MANUFACTURE OF ETHYL ALCOHOL FROM WOOD WASTE. U. S. Dept. Agr. Bul. 983, 100 p., illus.

(24) *Lipka, J.
1918. GRAPHICAL AND MECHANICAL COMPUTATIONS. 264 p., illus. New York.

(25) Marvin, C. F.
1924. A NEW PRINCIPLE IN THE ANALYSIS OF PERIODICITIES. U. S. Mo. Weather Rev. 52: 85–89, illus.

(26) ———
1924. FITTING STRAIGHT LINES TO DATA GREATLY SIMPLIFIED WITH APPLICATION TO SUN-SPOT EPOCHS. U. S. Mo. Weather Rev. 52: 89–91, illus.

(27) Merriman, M.
1910. METHOD OF LEAST SQUARES.

(28) *Mills, F. C.
[1924]. STATISTICAL METHODS APPLIED TO ECONOMICS AND BUSINESS. 604 p., illus. New York.

(29) ———
1924. THE MEASUREMENT OF CORRELATION AND THE PROBLEM OF ESTIMATION. Jour. Amer. Statis. Assoc. 19 (n. s. 147): [273]–300, illus.

(30) Miner, J. R.
1922. TABLES OF $\sqrt{1-r^2}$ AND $1-r^2$ FOR USE IN PARTIAL CORRELATION AND IN TRIGONOMETRY. 49 p. Baltimore.

(31) Neifeld, M. R.
1927. A STUDY OF SPURIOUS CORRELATION. Jour. Amer. Statis. Assoc. 22 (n. s. 159): 331–338.

(32) Ocagne, M. d'.
1899. TRAITÉ DE NOMOGRAPHIE. THÉORIE DES ABAQUES. APPLICATIONS PRATICUES. 427 p., illus. Paris.

(33) *Odell, C. W.
1926. THE INTERPRETATION OF THE PROBABLE ERROR AND THE COEFFICIENT OF CORRELATION. Ill. Univ. Bul. 32.

(34) *Pearl, R.
1923. INTRODUCTION TO MEDICAL BIOMETRY AND STATISTICS. 379 p., illus. Philadelphia and London.

(35) Pearson, K.
1924. TABLES FOR STATISTICIANS AND BIOMETRICIANS. Ed. 2, pt. 1. London.

(36) *Peddle, J. B.
THE CONSTRUCTION OF ALINEMENT CHARTS. Lefax Sheet 12–416. Philadelphia.

(37) ———
1910. THE CONSTRUCTION OF GRAPHICAL CHARTS. 109 p., illus. New York.

(38) Rietz, H. L., Carver, H. C., Crathorne, A. R., Crum, W. L., Glover, J. W., Huntington, E. V., and others.
1924. HANDBOOK OF MATHEMATICAL STATISTICS. 221 p., illus. Boston and New York.

(39) Secrist, H.
1917. AN INTRODUCTION TO STATISTICAL METHODS; A TEXTBOOK FOR STUDENTS, AND A MANUAL FOR STATISTICIANS AND BUSINESS EXECUTIVES. 482 p., illus. New York.

(40) ———
1920. STATISTICS IN BUSINESS; THEIR ANALYSIS, CHARTING AND USE. 137 p., illus. New York.

(41) Smith, B. B.
1925. THE ERROR IN ELIMINATING SECULAR TREND AND SEASONAL VARIATION BEFORE CORRELATING TIME SERIES. Jour. Amer. Statis. Assoc. 20 (n. s. 152): 543–545.

(42) Varga, S.
1928. AN EXPRESSION FOR THE ASYMETRICAL TENDENCY OF FREQUENCY DISTRIBUTIONS. Jour. Amer. Statis. Assoc. 239 (n. s. 161): 35–39.

(43) *Wallace, H. A., and Snedecor, G. W.
1925. CORRELATION AND MACHINE CALCULATIONS. Iowa Agr. Col. Off. Pub. v. 23, no. 35, 47 p.

(44) West, C. J.
1918. INTRODUCTION TO MATHEMATICAL STATISTICS. 150 p., illus. Columbus.

(45) *Wright, W, G.
1925. STATISTICAL METHODS IN FOREST-INVESTIGATION WORK. Canada Dept. Int., Forestry Branch Bul. 77, 36 p.

(46) ——— Robertson, W. M., and Mulloy, G. A.
1924. FOREST RESEARCH MANUAL. Canada Dept. Int., Forestry Branch 93 p., illus. [The formula on p. 47, $n=\frac{s^2}{S}$ should be $n=\left(\frac{s}{S}\right)^2$.]

(47) Yule, G. U.
1919. AN INTRODUCTION TO THE THEORY OF STATISTICS. Ed. 5, enl., 398 p., illus. London.

(48) ———
1926. WHY DO WE SOMETIMES GET NONSENSE-CORRELATIONS BETWEEN TIME SERIES?—A STUDY IN SAMPLING AND THE NATURE OF TIME-SERIES. Jour. Roy. Statis. Soc. 89 (pt. 1): 1–69, illus.

MIMEOGRAPHED MATERIAL

Smith, B. B.
1923. HANDBOOK OF STATISTICAL TERMS AND METHODS. 17 p. U. S. Dept. Agr., Bur. Agr. Econ.

———
1923. THE USE OF PUNCHED CARD TABULATING EQUIPMENT IN MULTIPLE CORRELATION PROBLEMS. U. S. Dept. Agr., Bur. Agr. Econ. [Rpt.]

———
1924. A POPULAR DISCUSSION OF MULTIPLE CORRELATION. WHAT IT IS, AND HOW IT IS DONE. 11 p. U. S. Dept. Agr., Bur. Agr. Econ. [Rpt.]

———
1926. CORRELATION THEORY AND METHOD APPLIED TO AGRICULTURAL RESEARCH. 102 p., Illus. U. S. Dept. Agr., Bur. Agr. Econ. [Rpt.]

STATISTICAL PERIODICALS

Biometrika. Cambridge, England, The University Press.

Journal of the American Statistical Association. Concord, N. H., The Rumford Press. (Prior to June, 1922, the journal was called The Quarterly Publications of the American Statistical Association.)

Journal of the Royal Statistical Society. London, published by the society.

Metron. Padua, Tipografia Industrie Grafiche Italiane Padova.

ORGANIZATION OF THE UNITED STATES DEPARTMENT OF AGRICULTURE WHEN THIS PUBLICATION WAS LAST PRINTED

Secretary of Agriculture	Arthur M. Hyde.
Assistant Secretary	R. W. Dunlap.
Director of Scientific Work	A. F. Woods.
Director of Regulatory Work	Walter G. Campbell.
Director of Extension Work	C. W. Warburton.
Director of Personnel and Business Administration.	W. W. Stockberger.
Director of Information	M. S. Eisenhower.
Solicitor	E. L. Marshall.
Weather Bureau	Charles F. Marvin, *Chief.*
Bureau of Animal Industry	John R. Mohler, *Chief.*
Bureau of Dairy Industry	O. E. Reed, *Chief.*
Bureau of Plant Industry	William A. Taylor, *Chief.*
Forest Service	R. Y. Stuart, *Chief.*
Bureau of Chemistry and Soils	H. G. Knight, *Chief.*
Bureau of Entomology	C. L. Marlatt, *Chief.*
Bureau of Biological Survey	Paul G. Redington, *Chief.*
Bureau of Public Roads	Thomas H. MacDonald, *Chief.*
Bureau of Agricultural Economics	Nils A. Olsen, *Chief.*
Bureau of Home Economics	Louise Stanley, *Chief.*
Plant Quarantine and Control Administration.	Lee A. Strong, *Chief.*
Grain Futures Administration	J. W. T. Duvel, *Chief.*
Food and Drug Administration	Walter G. Campbell, *Director of Regulatory Work, in Charge.*
Office of Experiment Stations	———————, *Chief.*
Office of Cooperative Extension Work	C. B. Smith, *Chief.*
Library	Claribel R. Barnett, *Librarian.*

This bulletin is a contribution from

Forest Service	R. Y. Stuart, *Chief.*
Branch of Research	Earle H. Clapp, *Assistant Forester, in Charge.*
Division of Silvics	E. N. Munns, *In Charge.*

U. S. GOVERNMENT PRINTING OFFICE: 1931

For sale by the Superintendent of Documents, Washington, D. C. - - - - - Price 15 cents